TARE: WEIGH A NEVER-POISONED TWIN BEFORE READING BACKDOOR-DEFENSE COSTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Backdoor-defense leaderboards print a clean-accuracy drop and read it as the cost of removal. The drop is measured on the poisoned victim alone, so it cannot separate removal from what the defense does to *any* model, and it inherits where the victim started. Where the victim starts is, for three of BackdoorBench's sixteen attacks, a configuration file: WaNet, BPP and Input-Aware ship a `MultiStepLR` whose first milestone is the epoch count, so their victims never anneal and are the lowest-accuracy victims in $30/31$ public CIFAR cells at $\leq 5\%$. On PreAct-ResNet18, fine-tuning-family defenses return a low start to a level of their own, so on those victims the published cost is negative, the benchmark's rating clips the "gain" to zero, and 2 of the 48 citing defense papers we read rest a no-cost claim on those cells; re-run with their released code, TSBD and CGD "gain" on a never-poisoned model too. We isolate the cause with a 2×2 that edits only that scheduler line, its swapped arms self-registered before they ran: the sign of the fine-tuning family's clean-model cost reverses both ways while its published gain on the annealed victim only shrinks toward zero, $44/44$ seeds following the schedule, replicated on BPP, FT-SAM, CIFAR-100 and VGG19-BN and induced in a second toolkit. The instrument is **TARE**: run the identical defense on a never-poisoned *twin* of the same recipe, schedule and seed (on BackdoorBench, ≤ 10 poisoned images, admitted only below 5% attack success); what the twin loses is the tare. On BackdoorBench's BadNets grid seven of the eight defenses charge the twin (Neural Cleanse only where its detector fires), $+0.13$ points (fine-tuning) to $+5.70$ (I-BAU); the eighth, ABL, destroys it. Within an attack the start cancels from rankings, so the tare re-orders nothing there; what poisoning adds beyond it is printed under two estimators and not corrected, its removal share being unidentified. We ship the three-key patch, a signed tare column over 7 attacks $\times$ 8 defenses of the CIFAR-10 roster, and TARE-Z, a twin-free estimator for seed-stable defenses.

1 INTRODUCTION

BackdoorBench scores a defense by two columns, the attack success left after it and the clean accuracy it costs; the second is measured on the poisoned victim alone and read as the price of removal. Where the victim starts is, for three of the benchmark's sixteen attacks, a configuration file: WaNet, BPP and Input-Aware ship a `MultiStepLR` whose first milestone is the epoch count, so their victims never anneal (Section 4.2), and on the public leaderboard they are the lowest-accuracy victims in $30/31$ CIFAR cells at rates $\leq 5\%$. For fine-tuning and NAD the accuracy *after* the defense is a level ($L_{\mathcal{D}}$, Section 3.2) of the defense's own (Figure 2): every never-annealed or self-damaged CIFAR-10 / PreAct-ResNet18 victim at 10%, LIRA's at 62.5% included, is returned to one 90.0–93.5% band while a healthy annealed one is not lifted (BadNets, starting at 91.32%, as low as WaNet, is lowered). So the published cost on such a victim is its start minus a constant that belongs to the defense, booked as a *gain* where the level lies above the start: the fine-tuning family "gains" on the three never-annealed victims and, in the single-seed leaderboard cells, on no healthy annealed one; the defense effectiveness rating's (DER) $\max(0, \cdot)$ clip zeroes it there; and two of the 48 defense papers we read in full carry a no-cost claim resting decisively on those cells (TSBD, CGD; a third, PDB, correlationally).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Edit that one `lr_scheduler` line and the clean-accuracy cost a defense charges a never-poisoned twin changes sign both ways (Figure 1a): $131/132$ (seed, defense) cells follow the schedule, not the attack. Within one attack every defense shares the victim, so by construction the victim's start cancels from the column's own rankings; it enters wherever the column is read alone—"improves over No Defense", DER's $\max(0, \cdot)$ clip (Section 4.4), cross-attack comparisons—and there we trace it. Run the identical defense on a never-poisoned twin and it, too, is charged (Table 1). Such a twin has been run before (Table 6), reported beside the cost, never subtracted from it.

Separating where the victim started from what poisoning adds to the bill needs a control that none of the six protocols we audit (Table 5) trains: a never-poisoned twin of the same recipe, schedule and seed, put through the identical defense. That is **TARE**, named for the weight a scale subtracts because it belongs to the container: what the twin loses is the tare, and what the poisoned arm loses beyond it is the poisoning-by-defense interaction, not trigger erasure (Section 3.1). The paired contrast differs from it by the pre-defense gap g, and which share of either is removal is not identified, so we print both beside the reported cost and correct nothing (Table 1). Our contributions:

- We find that three of BackdoorBench's sixteen attacks ship a `MultiStepLR` that never fires, and show with a 2×2 that edits only that line, its swapped arms registered, self-hosted, before they ran, that the tare's sign reverses both ways (Table 2, Figure 1a), replicated on BPP, FT-SAM, CIFAR-100, VGG19-BN and, induced, in a second toolkit, and at 5% poisoning (1% arms voided, predictions recorded failed; Appendix S35).
- We trace the artifact through the benchmark's rating into the papers that cite it: DER's clip hides the sign of these costs in $13/15$ never-annealed and $0/25$ healthy annealed single-seed leaderboard cells (descriptive, Appendix S39); TSBD and CGD, re-run with their code, "gain" on a never-poisoned twin (38% and 66% of the victims' gains), and dropping or re-referencing two cells flips TSBD's headline average; 2 of 48 defense papers read in full (one coder) rest a no-cost claim decisively on those cells.
- We propose TARE and measure its tare: seven of the eight defenses on BackdoorBench's BadNets grid charge a never-poisoned twin $+0.13$ to $+5.70$ points ($5/7$ at $0.05/7$, $7/7$ under Benjamini–Hochberg $q = 0.05$, Section 3.4's rule; Figure 1b), more than the whole reported cost for SAU, I-BAU; ABL destroys $41/50$ of Table 4's twins, a failure record; under a threshold rule that never sees attack success, ANP's tare rises from $+2.49$ to $+6.44$. We measure the accounting cost $=$ start $-$ level $+$ residual on twins and report its registered failures (Section 4.3).
- We ship the three-key patch, a signed tare column over 7 attacks $\times$ 8 defenses (re-ranking nothing within an attack, by construction; Table 4), TARE-Z, a twin-free estimator whose mean |error| (Eq. 6) is 0.12–0.68 points where a defense's level is seed-stable, and an archive that regenerates every number.

2 Related Work

Weight-space backdoor defenses and the score that ranks them. Weight-editing defenses—pruning (FP), fine-tuning (FT), masking (ANP), distillation (NAD), trigger inversion (NC), unlearning (ABL, I-BAU, SAU) and others (Liu et al., 2018a; Zhu et al., 2023a; Zhang et al., 2022; Wu & Wang, 2021; Li et al., 2021b; Wang et al., 2019; Li et al., 2021a; Zeng et al., 2022; Wei et al., 2023; Peng et al., 2025; Tong et al., 2025; Pawlak et al., 2025)—all report a clean-accuracy cost beside an attack-success drop, which the benchmark that scores them (Wu et al., 2025) folds into one rating (Zhu et al., 2023a).

Measurement audits of machine-learning benchmarks. This paper is a measurement audit and asks to be judged by that genre's standards, from tuned-baseline audits to *rankings* and *confounds* (Ferrari Dacrema et al., 2019; Musgrave et al., 2020; Schaeffer et al., 2023; Bouthillier et al., 2021; Bean et al., 2025; Singh et al., 2025; Qian et al., 2026; Gordienko et al., 2026; Neuhof & Benjamini, 2026; Su et al., 2025; Schaeffer et al., 2026; Harmon et al., 2025). Nasr et al. (2025) audit the *other* column of the same table. We follow Bouthillier et al. (2021)'s structure (estimand, ideal estimator, practice with one arm removed, bias quantified) with a same-recipe control arm.

The gain-score problem. The tuned-baseline audits above share the shape, not the mechanism: there the weak reference is a competing method run under a weaker recipe, and equalizing the recipe closes the gap; here the weak reference is the defense's own input, and equalizing the recipe only

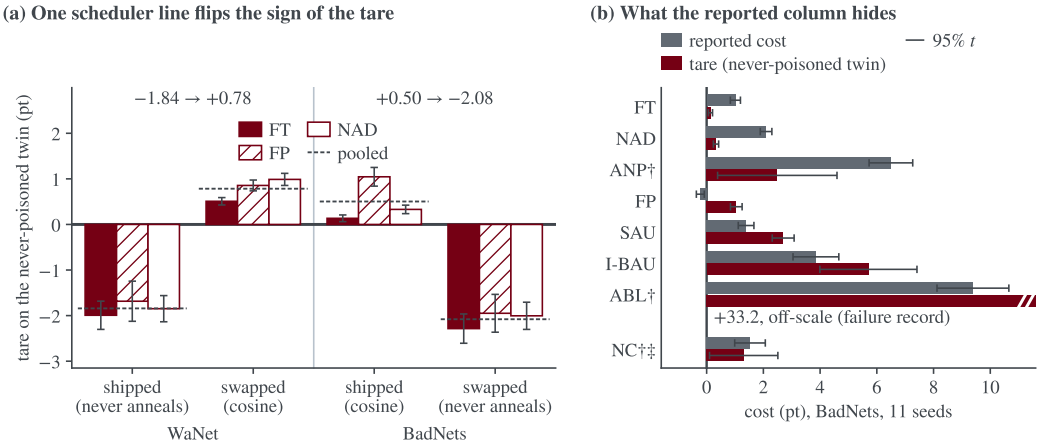


Figure 1: **(a) One `lr_scheduler` line reverses the tare's sign both ways** (FT, FP, NAD on never-poisoned twins; Table 2; dashed: pooled). **(b) What the reported column hides**: reported cost against tare, BadNets grid in Table 1's order (†, ‡ as there; ABL off-scale, a failure record). Whiskers: 95% t-intervals over seeds.

moves the cost toward zero (Section 4.2) because, at the benchmark's FT configuration, fine-tuning and NAD return every never-annealed or self-damaged CIFAR-10 victim to their level. That is the pretest–posttest gain-score problem (Lord, 1967; Cronbach & Furby, 1970): a treatment that resets its input makes before-minus-after the pretest minus a constant (Figure 2), and the reference decides the sign; Schaeffer et al. (2023) describe the nonlinear-metric artifact DER's $\max(0, \cdot)$ is.

Machine unlearning, which already ships the control we ask for. Unlearning evaluates removal against a from-scratch retrain (Maini et al., 2024; Shi et al., 2024; Zhang et al., 2024) and audits its own metrics (Kim et al., 2026; Cho et al., 2025; Khan et al., 2026), yet still reads utility drop against the *pre-unlearning* model (Su et al., 2026) (Appendix S23). **We claim no priority on the observation itself**—a systematization ran a no-attack arm and recommended this control (Abad et al., 2025), two defenses report the backdoor-free drop (Peng et al., 2025; Tong et al., 2025), and an LLM study attributes its capability drop to "the cost of continual pretraining itself" (Bouger et al., 2026)—**only on what is done with it (Table 6): an estimand with identification conditions, a same-recipe control in vision, the start gap's cause isolated and traced through the scoring function into citing papers, and a twin-free estimator with a stated domain of validity.**

3 Method

3.1 The Estimand and Its Two Estimators

Fix a capability metric Q (clean test accuracy) and an attack-success metric A. An *installation* ω collects everything the evaluator does not choose—base checkpoint, data order, poison placement, the defense's own stochasticity—and defines a poisoning run $\theta_{\text{pois}}(\omega)$ and its *clean twin* $\theta_{\text{twin}}(\omega)$, the identical run with each poisoned example replaced by its unmodified original, so that everything but the poison matches; both pass through the same defense $\mathcal{D}_b$ at the defender's clean-data budget b (Algorithm 1 and Figure 3, Appendix S7). On BackdoorBench, which refuses a zero rate, θ_{twin} is the attack's own pipeline at poisoning ratio (`pratio`) 0.0002 (ten images; one, at 0.00002, where ten install), same seed and schedule, admitted only if it passes A4's 5% gate; "never-poisoned" means this (a zero-image twin, guard bypassed, moves the tare by at most 0.35 points, one of nine intervals excluding zero; Appendix S34). Sharing the seed fixes recipe, schedule and data order, not the realized model, so pairing does not narrow the intervals (Appendix S46). The cost column is *named* for the extra capability the defense costs *because the model was poisoned*, holding the

defense fixed:

$$\tau_{\text{int}}(\mathcal{D}_b) \;=\; \mathbb{E}_\omega\Big[\big(Q(\theta_{\text{pois}})-Q(\mathcal{D}_b\theta_{\text{pois}})\big)-\big(Q(\theta_{\text{twin}})-Q(\mathcal{D}_b\theta_{\text{twin}})\big) \;\Big|\; A(\mathcal{D}_b\theta_{\text{pois}}(\omega)) \le \alpha\Big], \quad (1)$$

conditioned on removal, so the poisoned arm is one whose backdoor is gone (its rate in Appendix S3; the tare needs no such gate); it is the total effect of poisoning on the defense's bill. The price of erasure, $\tau_{\text{rem}} = \mathbb{E}_\omega[Q(\mathcal{D}_b\theta_{\text{pois}}) - Q(\mathcal{D}_b\theta^\circ)]$ with θ° a never-poisoned model at the poisoned arm's start, is not identified: θ° depends on how the start was lowered, and two routes give it opposite signs (Appendix S5). The literature reports the drop from a reporting baseline; the twin adds two estimators:

$$\text{reported (defense) cost} \quad c_{\text{def}} \;=\; Q(\theta_{\text{base}}) - Q(\mathcal{D}_b\theta_{\text{pois}}), \quad (2)$$

$$\text{interaction (DiD)} \quad c_{\text{int}}^{\text{DiD}} \;=\; \big[Q(\theta_{\text{pois}}) - Q(\mathcal{D}_b\theta_{\text{pois}})\big] - \big[Q(\theta_{\text{twin}}) - Q(\mathcal{D}_b\theta_{\text{twin}})\big], \quad (3)$$

$$\text{paired contrast} \quad c_{\text{int}} \;=\; Q(\mathcal{D}_b\theta_{\text{twin}}) - Q(\mathcal{D}_b\theta_{\text{pois}}) \;=\; c_{\text{int}}^{\text{DiD}} + g, \quad (4)$$

with $g = Q(\theta_{\text{twin}}) - Q(\theta_{\text{pois}})$ the pre-defense gap; the four cells are

	before, $Q(\theta)$	after, $Q(\mathcal{D}_b\theta)$	row difference
poisoned θ_{pois}	$Q(\theta_{\text{pois}})$	$Q(\mathcal{D}_b\theta_{\text{pois}})$	reported cost c_{def} (drift zero)
twin θ_{twin}	$Q(\theta_{\text{twin}})$	$Q(\mathcal{D}_b\theta_{\text{twin}})$	tare c_{clean}

the column differences being g before and the paired contrast after, and the difference of the two row differences the DiD. Eq. 3 identifies Eq. 1 under **(A1)** the arms differ only in the poison, **(A2)** $\mathcal{D}_b$ is the same map on both arms, **(A3)** removal is achieved, $A(\mathcal{D}_b\theta_{\text{pois}}) \le \alpha = 0.05$, and **(A4)** the twin is admissible, $A(\theta_{\text{twin}}) \approx 0$ under the same rule (the tare needs only A2 and A4; A2–A4 fail where Section 3.3 says). The two estimators are one identity apart, and where $g \neq 0$ which part of the paired contrast is removal is not identified: a starting gap alone produces an interaction of either sign, depending on how the gap was made (Appendix S5). **TARE prints the tare $c_{\text{clean}} = Q(\theta_{\text{twin}}) - Q(\mathcal{D}_b\theta_{\text{twin}})$ and the paired contrast beside c_{def}, and corrects nothing.**

3.2 What the Reported Cost Decomposes Into

Adding and subtracting the twin's two states splits the reported cost exactly,

$$c_{\text{def}} \;=\; \underbrace{c_{\text{int}}^{\text{DiD}}}_{\text{interaction}} \;+\; \underbrace{\big[Q(\theta_{\text{twin}}) - Q(\mathcal{D}_b\theta_{\text{twin}})\big]}_{\text{tare } c_{\text{clean}}} \;+\; \underbrace{\big[Q(\theta_{\text{base}}) - Q(\theta_{\text{pois}})\big]}_{\text{reporting-baseline drift}}, \quad (5)$$

the drift being zero in every headline here ($\theta_{\text{base}} = \theta_{\text{pois}}$). **The fixed-point accounting (Figure 2).** For a defense whose defended level $L_{\mathcal{D}}$ does not depend on where the model started, $Q(\mathcal{D}_b\theta) = L_{\mathcal{D}} + \varepsilon(\theta)$ on either arm, $L_{\mathcal{D}}$ its fixed point, so $c_{\text{clean}} = Q(\theta_{\text{twin}}) - L_{\mathcal{D}} - \varepsilon(\theta_{\text{twin}})$, $c_{\text{def}} = Q(\theta_{\text{pois}}) - L_{\mathcal{D}} - \varepsilon(\theta_{\text{pois}})$ and $c_{\text{int}}^{\text{DiD}} = -g + [\varepsilon(\theta_{\text{twin}}) - \varepsilon(\theta_{\text{pois}})] = -g + c_{\text{int}}$: the reported cost is the victim's start minus a constant that belongs to the defense plus what the poison leaves at that level, and the interaction is the start gap returned plus that residual. Whether a defense has such a level is measured, not assumed: on PreAct-ResNet18 it holds for FT, NAD, FT-SAM and SAU, not for I-BAU, VGG19-BN or a damaged start (Section 4.3, Figure 10, Appendix S48); across four CIFAR-10 cells whose twins span 3.07 points, the defended twin moves under a point for FT, NAD, FP, SAU and FT-SAM and 3.26 for I-BAU (Table 36).

3.3 Admissibility and Scope

A3 needs a utility floor, since a residual-attack-success gate alone is satisfiable by destroying the model (Appendix S3). **A4.** A twin is a control only if it is clean *by the metric that scores the defense*: the admissible CIFAR-10 twins score 0.6–4.5% residual attack success under A4's 5% gate; a ten-image blended twin scores 11–16% against a 2–3% never-exposed floor and does *not* qualify (a one-image twin does, Appendix S20); SIG's and Trojan's twins fail it and their cells are descriptive (Appendix S52). **A2** holds at a stated level, policy or realized repair: a conditional defense—repair only when the detector fires—is one function at the policy level, so its twin arm is well defined and a no-op on the twin zeroes the tare; forcing the repair branch on a never-poisoned twin costs $+3.40$ $[+3.21, +3.59]$ points (Appendix S26). **Manufactured cost.** On a model with no backdoor the reported cost is the pure tare, $c_{\text{def}}^0(b) = Q(\theta_{\text{twin}}) - Q(\mathcal{D}_b\theta_{\text{twin}})$ with $\tau_{\text{int}} \equiv 0$ by construction,

and the budget b moves it: fine-pruning costs 3.79 $[3.25, 4.33]$ points at 1% clean data and 0.28 $[0.09, 0.47]$ at 100% (Appendix S16); a cost column is not comparable across papers that omit b. **Gating, once.** Table 1 averages every column of a row over that defense's A3-passing seeds, so its identity holds within the row; Table 4 prints the tare unconditional and, with Section 4.2, gates poisoned-arm quantities on A0 (pre-defense attack success ≥ 0.90) and prints A3 beside them; A3 selects on one arm's post-treatment outcome, its effect on the residue measured in Appendix S3.

3.4 TARE-Z, Replicates and Intervals

The twin requires knowing which examples were poisoned. TARE-Z replaces it by an independently trained clean model of the same architecture and recipe,

$$\hat{c}_{\text{int}} = Q(\mathcal{D}_b\theta_{\text{surr}}) - Q(\mathcal{D}_b\theta_{\text{pois}}), \qquad \hat{c}_{\text{int}} - c_{\text{int}} \;=\; Q(\mathcal{D}_b\theta_{\text{surr}}) - Q(\mathcal{D}_b\theta_{\text{twin}}), \tag{6}$$

so its error is *exactly* a clean-versus-clean difference, mean-zero by exchangeability, before conditioning on A3; its per-use size is the seed-to-seed spread of the defense's level (Section 4.7), and only a surrogate of another recipe can be biased. **The replicate is the installation**, an end-to-end re-run of poisoning, twin and defenses (Bouthillier et al., 2021); two never-poisoned models differing only in seed ($\tau_{\text{int}} \equiv 0$) calibrate the instrument (Appendix S30). Cost estimates carry 95% t-intervals over installations; rank statistics use an installation cluster bootstrap with rank confidence intervals (Neuhof & Benjamini, 2026; Klein et al., 2018), a pair being separable when its sign survives Benjamini–Hochberg at $q = 0.05$ (Appendix S19); a null is reported as the interval itself, never as a proven zero.

4 Experiments

4.1 Setup

Vision. BackdoorBench grids run the framework's own pipeline (PreAct-ResNet-18, CIFAR-10, its 10% rate, 100 recovery epochs) with the twin of Section 3.1 (Appendix S20). Our own CIFAR-10 grid trains ResNet-18 with a BadNets trigger at 5%, twin at 0%, and runs fine-tuning, fine-pruning, ANP, NAD and NC (Wu & Wang, 2021; Li et al., 2021b; Wang et al., 2019) at the standard 5% clean-data budget (Appendix S12). At the framework's FT/NAD operating point this rig leaves residual attack success 0.37/0.48 on BadNets against the framework's 0.015/0.014 (fine-tuning clears $1/8$ installations against $11/11$), so the two rigs are not compared in level (Table 14). Every run here is CIFAR-10/100 on ResNet-18, PreAct-ResNet18, VGG19-BN or BackdoorBox's ResNet-18; other datasets and backbones enter only by leaderboard reading (Table 3). An exploratory language-model check at budget points that are not deployable operating points is Appendix S2; no claim rests on it. θ_{base} in Eq. 2 is the pre-defense poisoned model in both modalities; one defense policy runs on both arms (Section 3.3).

4.2 The Scheduler Block, Manipulated: the Main Result

The cause is three configuration keys, the scheduler block. BackdoorBench's prototype config, the base every attack inherits, trains every victim under `CosineAnnealingLR` for 100 epochs; the benchmark's `default.yaml` for WaNet (Nguyen & Tran, 2021) and BPP (Wang et al., 2022b) override it with `MultiStepLR` at milestones $[100, 200, 300, 400]$, and its Input-Aware (Nguyen & Tran, 2020) schedule carries the same milestones, so the first decay never comes and the victim trains at lr 0.01 throughout; the attacks' own releases place these milestones inside 600- to 1000-epoch runs, and the benchmark cut the run and kept them (Appendices S21, S33). BackdoorBench's 2022 paper exempts training-controllable attacks from its cosine default and lists these milestones for WaNet and Input-Aware (its Table 4, archived in the supplement): a defender handed such a model does gain from fine-tuning, and c_{def} is that defender's bill, but the gain is not removal and is not comparable across attacks. **A 2×2 changes only that**, its swapped arms registered in a dated, self-hosted file before they ran (the shipped arms were known; Appendix S43). With the two schedulers crossed over WaNet and BadNets (Gu et al., 2017), the undefended twins move from 91.50 to 94.18% (WaNet) and from 93.71 to 91.09% (BadNets), and the tare reverses with the scheduler, not the attack: -1.84 $[-2.18,-1.50]\rightarrow+0.78$ $[+0.70,+0.87]$ on WaNet, $+0.50$ $[+0.42,+0.59]\rightarrow-2.08$ $[-2.40,-1.76]$ on BadNets, twelve per-defense intervals excluding zero, $131/132$ cells and $44/44$ seeds (three

Table 1: **Per-defense accounting on BackdoorBench's BadNets grid (CIFAR-10, eleven installation seeds), in points.** Reported cost c_{def} against the pre-defense poisoned model; tare c_{clean}, the twin's loss to the same defense; paired contrast (Eq. 4) and interaction (DiD, Eq. 3), differing by $g = +2.06$ [+1.90,+2.23], no share of which is identified as removal; A3: seeds clearing the backdoor (residual ASR ≤ 0.05) over seeds attempted; rows average those seeds. Brackets: 95% t-intervals. †: A2 at the policy level (ANP's threshold is picked by test-set attack success, an oracle) or a failure record (ABL); ‡: NC over the seeds clearing A3, an untouched twin entering at zero (Appendix S19). Pooled row: both signs; our grid: Table 25.

Defense	reported c_{def} (pt)	tare c_{clean} (pt)	paired $Q(\mathcal{D}\theta_{\mathrm{twin}}) - Q(\mathcal{D}\theta_{\mathrm{pois}})$ (pt)	interaction, DiD (pt)	A3 (k/n)
fine-tuning	+1.01 [+0.84,+1.19]	+0.13 [+0.06,+0.21]	+2.94 [+2.75,+3.14]	+0.88 [+0.68,+1.08]	11/11
NAD	+2.10 [+1.89,+2.30]	+0.33 [+0.24,+0.42]	+3.83 [+3.59,+4.07]	+1.77 [+1.52,+2.02]	11/11
ANP†	+6.50 [+5.73,+7.27]	+2.49 [+0.39,+4.60]	+6.07 [+3.75,+8.39]	+4.01 [+1.69,+6.32]	11/11
fine-pruning	−0.22 [−0.36,−0.08]	+1.05 [+0.84,+1.25]	+0.80 [+0.50,+1.09]	−1.27 [−1.60,−0.93]	11/11
SAU	+1.39 [+1.11,+1.67]	+2.70 [+2.31,+3.08]	+0.76 [+0.26,+1.25]	−1.31 [−1.80,−0.82]	11/11
I-BAU	+3.85 [+3.05,+4.66]	+5.70 [+3.99,+7.42]	+0.21 [−1.84,+2.27]	−1.85 [−3.93,+0.23]	11/11
ABL†	+9.39 [+8.12,+10.7]	+33.2 [+29.1,+37.3]	−21.8 [−26.3,−17.2]	−23.8 [−28.3,−19.3]	11/11
Neural Cleanse†‡	+1.53 [+0.99,+2.07]	+1.31 [+0.10,+2.51]	+2.28 [+1.16,+3.39]	+0.22 [−0.85,+1.29]	8/11
pooled, 7 excl. ABL	+2.34 [+2.10,+2.57]	+1.98 [+1.50,+2.46]	+2.42 [+1.81,+3.04]	+0.36 [−0.26,+0.98]	74/77

Table 2: **The 2×2 scheduler crossover: only the `lr_scheduler` line differs between a shipped row and its swapped counterpart.** Tare per defense (pt, unconditional), mean ± half-width of the 95% t-interval over installation seeds; "pooled": the three defenses averaged within a seed; "twin": the undefended twin's accuracy, the registered prediction's quantity; "cells": (seed, defense) cells whose sign follows the scheduler. Published cost, DiD and A3 per row: Appendix S21.

attack	victim's scheduler	seeds	fine-tuning	fine-pruning	NAD	pooled	twin (%)	cells
WaNet	step, never fires (shipped)	11	−1.99 ±0.31	−1.68 ±0.44	−1.85 ±0.29	−1.84 ±0.34	91.50	33/33
WaNet	cosine (swapped)	11	+0.51 ±0.08	+0.86 ±0.12	+0.99 ±0.13	+0.78 ±0.09	94.18	33/33
BadNets	cosine (shipped)	11	+0.13 ±0.07	+1.05 ±0.20	+0.33 ±0.09	+0.50 ±0.09	93.71	32/33
BadNets	step, never fires (swapped)	11	−2.29 ±0.32	−1.95 ±0.41	−2.00 ±0.30	−2.08 ±0.32	91.09	33/33

defenses pooled) agreeing in sign (Table 2, Figure 1a; six defenses in Table 36). The patch improves the attack too: WaNet's victim rises from 90.05 to 92.69% and installs in 11/11 seeds against 8/11 shipped. The sign was the registered expectation (priors 0.70–0.85, Appendix S21); the crossover adds attribution—the reversal follows the schedule, not the attack—and size. Over installing seeds (A0; A3 21/24 and 33/33) the published cost moves toward zero (−3.07 [−3.42,−2.73] to −0.25 [−0.66,+0.16]) while the DiD moves from −1.30 [−2.00,−0.61] to −1.03 [−1.49,−0.58] (not distinguishable; no equivalence bound claimed): the start gap returned, not identified removal (Appendix S5). The schedule moves removal too: without the anneal FT, FP and NAD clear BadNets in 15/33 (seed, defense) cells against 33/33 with it (descriptive; Appendix S21). At 5% poisoning the crossover replicates (registered; six intervals excluding zero); at 1% the victims mostly fail to install, so those arms are void and all four registered predictions count as failed (Appendix S35). With all eight defenses on both arms (registered; Appendix S37), 3/28 pairs of the benchmark's own WaNet DER order change with the patch (FT/NAD, FP/I-BAU, FP/SAU), none separable seed by seed on the shipped victim (0/3; Table 43).

4.3 The Sign Replicates; the Level Does Not Always

The sign replicates. On BPP the tare moves from −2.01 [−2.63,−1.40] to +0.87 [+0.69,+1.06] (Input-Aware has no admissible twin, A4 0/5); FT-SAM flips as fine-tuning does (−2.14 [−2.46,−1.81]→+1.09 [+0.92,+1.26]), SAU lands at a lower level of its own (91.4%) and I-BAU has none (Figure 10); on CIFAR-100 the twin's rise doubles (+6.01 on WaNet, +6.75 on BPP) and an annealed control given the block flips the other way (Blended, +2.76 [+2.50,+3.03]→−5.23 [−7.60,−2.86]); on VGG19-BN fine-tuning flips (−2.30 [−2.72,−1.89]→+2.09 [+1.59,+2.59]); and in a second toolkit, BackdoorBox, whose shipped 200-epoch WaNet run anneals, cutting it to 100 epochs under its own milestones starts the twin 8.61 points lower and hands fine-tuning and NAD a "gain" (−6.10, −7.23) that decay erases or reverses (+0.01, +0.81), while pruning, which trains nothing, charges both (Appendix S33). **The**

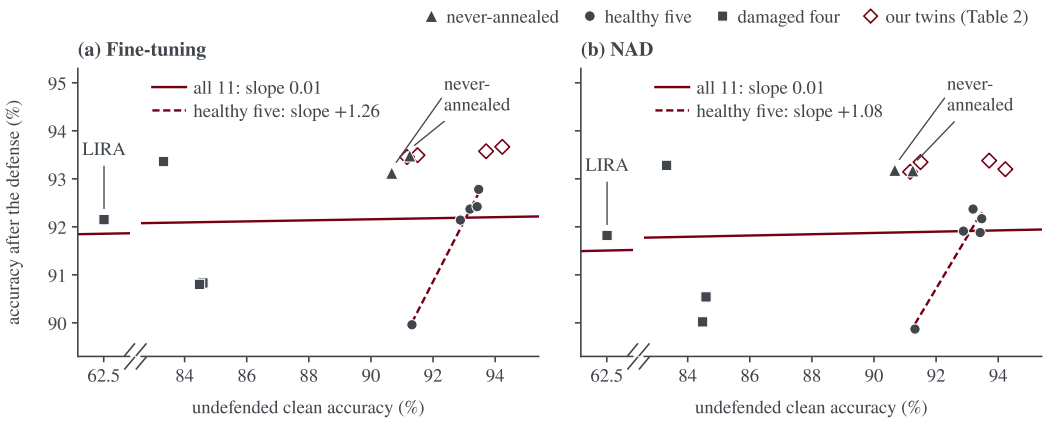


Figure 2: **On PreAct-ResNet18, FT (a) and NAD (b) return a never-annealed or self-damaged start to a level of their own**: after against before, public CIFAR-10 leaderboard at 10% (single-seed cells, grouped after reading the leaderboard; slopes descriptive, 0.06 and 0.07 without LIRA; start alone does not set the after: BadNets starts as low as WaNet and FT lowers it; hollow: our twins; Appendices S38, S39).

Table 3: **The never-annealed victims start lowest across the public leaderboard's datasets and backbones.** Other attacks' mean clean accuracy minus that of WaNet, BPP and Input-Aware, over the rates $\leq 5\%$ with public entries; last column: cells where the three are the three lowest (sources: Appendix S38). †: on Tiny-ImageNet the three anneal once, at epoch 100, a different contrast; GTSRB is saturated.

dataset	PreAct-R18	VGG19-BN	ConvNeXt-T	ViT-B/16	lowest three
CIFAR-10	+2.1	+2.6	+4.1	+4.9	15/16
CIFAR-100	+6.3	+5.9	+9.6	+10.0	15/15
Tiny-ImageNet†	—	−2.4	+10.5	+11.3	4/10
GTSRB (saturated)	+0.3	+0.9	+0.4	−7.3	2/16

level does not always: fine-pruning's fixed point fails on CIFAR-100, VGG19-BN's defended twins stay 1.72 points apart, and three further ways to lower a start—SIG's self-damage, a 30-epoch run, a 20% dose—reproduce the sign but not the level (registered fixed-point predictions, failed; Appendices S49–S51). **The level is the defense's, not the tail of a cosine:** 76% of the shipped WaNet twin's rise arrives before any decay (Appendix S47). The victim's schedule sets the start; the defense's configuration sets where a low start lands. More poison lowers the start too: from 5% to 20% BadNets every defense's published cost falls (8/8; a rise was registered); registered for Blended and LF, where the start moves by +1.02 and +0.50, it reversed (0/4 pairs fall; a failed registered prediction), while the tare, one twin per seed, cannot move (Appendix S51). Each replication's protocol was written, self-hosted, before its new runs; six analysis decisions came after data, one (an SAU exclusion) after its outcome (Appendix S43).

4.4 What the Artifact Reached: Leaderboard, Rating, Papers

BackdoorBench's public CIFAR-10 / PreAct-ResNet18 leaderboard prints, for every attack at five poisoning rates, the victim's clean accuracy and the accuracy after each defense (Figure 2). For fine-tuning and NAD the after is a level of the defense's own: victims spanning 31.0 points before are returned to a 3.5-point band (slopes 0.01 and 0.01 over all eleven attacks, 0.06 and 0.07 without LIRA, 0.18 and 0.22 on the eight all three defenses cover), while on five healthy annealed victims (BadNets, Blended, LF, SSBA, Trojan; grouped after reading the leaderboard, any cutoff between 84.6 and 91.3% selects the same five) the slopes are +1.26 and +1.08 (descriptive; they lean on BadNets), over every annealed attack, damaged ones included, +0.00 (Appendix S39): the defense replaces a low start and leaves a healthy one near where it was. The never-annealed attacks start lowest (Table 3). **The bias reaches the rating.** BackdoorBench's defense effectiveness rating over

Table 4: **The tare per (attack, defense) cell on CIFAR-10 / PreAct-ResNet18 at 10%, wherever an admissible twin exists.** Line 1 per defense: tare (pt), mean $\pm$ 95% t-half-width over seeds; line 2 (*paired*): $Q(\mathcal{D}\theta_{\text{twin}}) - Q(\mathcal{D}\theta_{\text{pois}})$ over installing seeds (A0: attack success ≥ 0.90), DiD plus the pre-defense gap (Eq. 4). Twin construction, SIG's and Trojan's descriptive cells, published costs, DER, seeds, five more defenses: Appendix S52; patched WaNet/BPP: Table 36. †: A2 at the policy level only (ANP's threshold; NC's detector; ABL); NC: twins whose detector fired | cost given firing. ‡: under half the poisoned cells clear A3 (FT, FP, NAD: 1 of 69 on Blended, LF, FTrojanNN, Refool); entry: a post-defense contrast. $^{\natural}$: ABL destroys $41/50$ of these twins; its row is a failure record, off-scale, outside every sign count and the ordering. Intervals per-cell, nominal: of 49 admissible cells with intervals (ABL in, NC out), 44 exclude zero at 0.05 and 15 at Bonferroni $0.05/K$, $K = 63$ over the nine-attack table (Appendix S52). Columns: attacks; A4: twins passing the 5% gate over seeds.

defense	BadNets	Blended	LF	FTrojanNN	Refool	WaNet	BPP
victim schedule	cosine	cosine	cosine	cosine	cosine	never	never
undefended twin	93.71	93.84	93.92	93.84	93.74	91.50	91.48
A4: clean twins	$11/11$	$8/8$	$5/5$	$5/5$	$5/5$	$11/11$	$5/5$
FT	$+0.13 \pm 0.07$	$+0.20 \pm 0.11$	$+0.11 \pm 0.14$	$+0.24 \pm 0.15$	$+0.16 \pm 0.15$	-1.99 ± 0.31	-2.07 ± 0.75
paired	$+2.94$	$+0.86^{\ddagger}$	$+1.32^{\ddagger}$	$+0.84^{\ddagger}$	$+1.69^{\ddagger}$	$+0.30$	$+0.22$
FP	$+1.05 \pm 0.20$	$+1.12 \pm 0.23$	$+1.08 \pm 0.46$	$+1.04 \pm 0.36$	$+0.91 \pm 0.30$	-1.68 ± 0.44	-1.82 ± 1.07
paired	$+0.80$	$+0.28^{\ddagger}$	$+0.39^{\ddagger}$	$+0.15^{\ddagger}$	$+0.84^{\ddagger}$	-0.06	$+0.25^{\ddagger}$
NAD	$+0.33 \pm 0.09$	$+0.36 \pm 0.15$	$+0.22 \pm 0.18$	$+0.47 \pm 0.25$	$+0.36 \pm 0.16$	-1.85 ± 0.29	-1.84 ± 0.64
paired	$+3.83$	$+1.65^{\ddagger}$	$+1.96^{\ddagger}$	$+1.27^{\ddagger}$	$+2.00^{\ddagger}$	$+0.37$	$+0.33$
NC†	$5/11 \mid +2.71$	$3/8 \mid +3.06$	$1/5 \mid +3.22$	$3/5 \mid +2.86$	$1/5 \mid +2.79$	$3/11 \mid -1.30$	$0/5 \mid —$
paired	$+1.94$	$+0.57^{\ddagger}$	$+2.35$	$+0.57$	$+1.12^{\ddagger}$	$+0.46$	-1.22
ANP†	$+2.49 \pm 2.10$	$+3.86 \pm 2.86$	$+3.25 \pm 3.08$	$+1.43 \pm 1.01$	$+4.72 \pm 3.62$	$+1.78 \pm 1.26$	$+4.28 \pm 2.48$
paired	$+6.07$	$+3.61$	$+4.09$	$+5.80$	$+3.28$	$+4.95$	$+2.51$
ABL$^{\dagger\natural}$	$+33.20 \pm 4.06$	$+27.83 \pm 11.98$	$+31.28 \pm 15.75$	$+32.36 \pm 12.73$	$+35.91 \pm 11.28$	$+27.77 \pm 6.49$	$+29.83 \pm 12.18$
paired	-21.75	-7.08	$-6.32^{\ddagger}$	-15.48	$+4.90^{\ddagger}$	$-6.27^{\ddagger}$	$-2.05^{\ddagger}$
I-BAU	$+5.70 \pm 1.71$	$+4.95 \pm 1.02$	$+5.98 \pm 2.26$	$+6.06 \pm 2.37$	$+5.66 \pm 1.68$	$+0.23 \pm 0.53$	$+0.71 \pm 1.67$
paired	$+0.21$	-0.42	$-0.61^{\ddagger}$	$-1.32^{\ddagger}$	$+0.26^{\ddagger}$	$+0.20$	$+0.06$
SAU	$+2.70 \pm 0.39$	$+3.06 \pm 1.76$	$+2.83 \pm 0.80$	$+2.92 \pm 0.51$	$+3.71 \pm 1.95$	-0.15 ± 0.62	-0.60 ± 0.96
paired	$+0.76$	$+1.50$	$+0.25$	-0.35	$-0.28^{\ddagger}$	-0.10	$+0.56$

11,492 cells (Zhu et al., 2023a; Wu et al., 2025) is

$$\text{DER} \;=\; \tfrac{1}{2}\big[\max(0, \Delta A) - \max(0, \Delta Q) + 1\big], \tag{7}$$

ΔA, ΔQ: attack-success and clean-accuracy drops. The fine-tuning family's published cost is negative on the never-annealed victims and positive on the healthy annealed ones, and the clip hides the negatives: computed from the site's accuracies (it prints no DER; FT, FP, NAD, NC, I-BAU; descriptive), it fires in $13/15$ never-annealed cells and $0/25$ healthy annealed ones (Appendix S39; seed-level counts on our grids, Appendix S52). **The papers.** Of the 252 papers citing either BackdoorBench paper, we read in full (one coder, no second rating; 11 journal-only papers unobtained, Appendix S54) the 48 that propose a defense against installed backdoors and asked whether each makes a clean-accuracy claim that depends on a WaNet, BPP or Input-Aware cell. 13 reuse the benchmark's victims and reproduce the artifact, unremarked, in at least 7 baseline columns; two claims rest on it causally, a third correlationally. *TSBD* (Lin et al., 2024) opens its CIFAR-10 results with an average "higher than the original No Defense models" (91.70 vs. 91.34%). Rerun with its code in the WaNet 2×2 (Appendix S53), it reproduces the gain on the shipped victim and not on the annealed one; on the never-poisoned twin it "gains" -1.11 $[-1.86, -0.35]$ shipped (38% of the victim's gain at its printed default; Table 59 sweeps the ratio) and costs $+2.59$ $[+1.37, +3.82]$ under cosine. *CGD* (Xu et al., 2025) reports "improving CA over the No Defense baseline by 3.4% for WaNet"; in the same 2×2 the twin gains -2.43 $[-3.00, -1.86]$ (66% of the victim's gain) and loses $+0.41$ $[+0.26, +0.57]$ under cosine. *PDB* (Wei et al., 2024) has a positive Tiny-ImageNet average only because of WaNet and BPP (correlational). Dropping or re-referencing its two never-annealed cells flips TSBD's average (Appendix S54); none of the 48 notes that the three victims are under-trained, and what rests on them is a sentence per method.

4.5 A Signed Tare Column for the 10% Leaderboard

Table 4 is the deliverable table (7 attacks, 118 GPU-hours; rows, columns and gates in its caption); it corrects no cell, and sign counts and the ordering leave ABL out (a failure record, not a cost). On the annealed victims every tare is positive ($35/35$ cells; on the four annealed victims besides BadNets

FT, FP and NAD remove these backdoors in 1 of 69 poisoned cells, so their paired entries there price nothing). On the two never-annealed victims FT, FP and NAD "gain" in $6/6$ cells and ANP† charges the twin in $2/2$, every nominal interval excluding zero ($3/8$ at the Bonferroni level); SAU, at a level of its own, and I-BAU, with none, have intervals covering zero in $4/4$ of their cells; where Neural Cleanse's detector fires ($3/11$ WaNet twins), its repair, a fine-tune, "gains" -1.30 as fine-tuning does. This pattern was read off the table after a registered prediction failed on BPP (Appendix S43). Ordering these eight defenses by mean tare over the admissible attacks gives FT $<$ NAD $<$ FP $<$ NC $<$ SAU $<$ ANP $<$ I-BAU (ABL off-scale, last), NC at its policy-level tare with non-firing twins at zero (Appendix S26; under a deployer-side threshold rule that never sees attack success ANP's tare rises from $+2.49$ to $+6.44$ on BadNets and from $+1.78$ to $+6.79$ on WaNet—registered, two of three predictions held, the failed one that the two rules agree within a point; Appendix S36). A prototype-referenced DER (the same seed's BadNets twin as reference) re-ranks $0/28$ pairs over all nine attacks (Table 57): attribution changes, order does not, by construction. Five further defenses re-rank $1/78$ pairs (Appendix S52).

4.6 What the Twin Charges per Defense, and What Poisoning Adds

At a fixed attack the charge is a property of the defense; at a fixed defense, of the victim's schedule. On BackdoorBench's shipped framework all eight BadNets clean-arm intervals exclude zero (nominal; among the seven besides ABL, $5/7$ at $0.05/7$ and $7/7$ under BH $q = 0.05$; Figure 1b, Figure 6, Appendix S19), and the charge runs, over the six reporting a positive cost (ABL aside), from 13% of what the defense reports (fine-tuning, $+0.13$ of $+1.01$; $+0.07$ $[-0.05, +0.20]$ on a zero-image twin) to 194% (SAU, $+2.70$ of $+1.39$); ABL, off-scale at 354%, collapses its twins (Table 30). **What poisoning adds is estimator- and defense-dependent, and both estimators are printed.** A reported $+2.34$ leaves $+0.36$ $[-0.26,+0.98]$ by difference-in-differences and $+2.42$ $[+1.81,+3.04]$ paired ($g = +2.06$), the per-defense DiD carrying both signs (Table 1; our own rig: Table 25). Read as $c_{\text{int}} - \beta g$ over $\beta = 0$ (paired), 1 (DiD) and 0.45, 1.2 (two start-lowering routes, two-point fits on our ResNet-18 rig; Appendix S5), the interaction keeps its sign for FT, NAD and ANP, not for NC, FP, SAU and I-BAU; which β is removal's is unidentified, so nothing is corrected. On a checkpoint carrying a backdoor with probability p, a defense costs $(1-p)\,c_{\text{clean}} + p\,c_{\text{def}}$ in expectation; benchmarks print the $p = 1$ corner; post hoc (the registered criterion failed), four BadNets pairs reverse below $p = 0.24$ and none survives on WaNet or Blended (Appendix S19).

4.7 What the Twin Costs to Run

The twin is not needed where a defense's clean-arm level is stable across seeds. Substituting a same-recipe clean model for the twin (Eq. 6) over 8 configurations of FT, FP, NAD and ANP on 4 installations is off by -0.02 $[-0.17,+0.12]$ points paired in the mean (Appendix S28); leaving one twin out of BackdoorBench's eleven-seed BadNets grid ($+0.00$ $[-0.51,+0.51]$ paired, excluding ABL) measures how reproducibly a level recurs. For fine-tuning, NAD, fine-pruning, SAU its per-use error (mean |error| 0.12–0.68) resolves the fine-tuning family's never-annealed sign, not a tare as small as fine-tuning's $+0.13$ on annealed victims; for Neural Cleanse, I-BAU, ANP, ABL (1.58–6.79) the defense's own randomness moves the level by more than one surrogate resolves, so seeds must be pooled. Across architectures it is ordinal only (Appendix S27). **Ex ante, without a twin: the defense's spread over two or three same-recipe clean models is the per-use error (Eq. 6); above the tare in question, pool seeds.**

5 Conclusion

We show that the clean-accuracy cost backdoor-defense leaderboards print is measured on the poisoned victim alone and inherits where that victim started. On BackdoorBench three of sixteen attacks ship a scheduler that never anneals; one `lr_scheduler` line reverses the sign of the fine-tuning family's cost on a never-poisoned twin, the benchmark's rating clips the resulting "gain" to zero, and two citing papers rest a no-cost claim on it. We propose TARE, the identical defense run on a never-poisoned twin, whose loss—the tare—is $+0.13$ to $+5.70$ points on BackdoorBench's BadNets grid (ABL aside). Patched, WaNet's victim anneals and its tare turns positive; printed signed and unclipped beside the reported column, the tare changes attribution and the cross-attack averages, not the within-attack order.

AI USE STATEMENT

The authors used an AI coding assistant to help implement experiment harnesses, orchestrate runs, draft and revise text, and critique drafts in simulated review rounds (the "panels" some supplement files mention). All experimental designs, claims, and their verification were determined and checked by the authors; all reported numbers were produced by the authors' code and inspected by the authors. No AI-generated research claim was included without independent verification by the authors.

ETHICS STATEMENT

This work installs backdoors only in models we train ourselves, on public datasets, and releases no new attack: every attack we use (BackdoorBench's BadNets (Gu et al., 2017), Blended (Chen et al., 2017), WaNet (Nguyen & Tran, 2021), BPP (Wang et al., 2022b), Input-Aware (Nguyen & Tran, 2020), SIG (Barni et al., 2019), LF (Zeng et al., 2021), Trojan (Liu et al., 2018b), FTrojanNN (Wang et al., 2022a) and Refool (Liu et al., 2020); LIRA (Doan et al., 2021) and SSBA (Li et al., 2021c) on the public leaderboard only; and the instruction-tuning poison of Souly et al. (2025)) is already published, and no poisoned weights are distributed. The paper's finding is an accounting correction and should not be read as a claim that deployed defenses are safe: we show that the clean-accuracy figure printed beside a defense does not isolate poisoning-conditioned cost from generic clean-arm cost, and that the relative contribution varies across defenses and settings, which changes what portion of the reported utility loss is attributable to removal but says nothing about whether removal succeeded or whether the defense is safe. Sections 3.1 and 3.3 state the conditions under which our estimand is identified, and we report the settings in which it is not.

REPRODUCIBILITY STATEMENT

The defenses are fine-tuning, fine-pruning (Liu et al., 2018a), NAD (Li et al., 2021b), Neural Cleanse (Wang et al., 2019), ANP (Wu & Wang, 2021), ABL (Li et al., 2021a), I-BAU (Zeng et al., 2022), SAU (Wei et al., 2023), FT-SAM (Zhu et al., 2023a), CLP (Zheng et al., 2022), MCR (Zhao et al., 2020), RNP (Li et al., 2023), NPD (Zhu et al., 2023b), D-BR (Chen et al., 2022), TSBD (Lin et al., 2024) and CGD (Xu et al., 2025), as implemented in BackdoorBench (Wu et al., 2025), in the methods' released code (TSBD, CGD) or in our harness (Appendix S12). The three keys were read at BackdoorBench commit `f02e353` (2025-07-04). The BackdoorBench-framework grids used at least 328 logged GPU-hours, the BackdoorBox runs 21.7, the zero-image twins 5.9, the rate crossover 15.1 and the eight-defense crossover 27.1(the ANP re-reading of Appendix S36 trained nothing). We report twenty-five self-hosted pre-registrations (no external registry; timing evidence and the decisions taken after data existed are in Appendix S43), failures reported as failed. The paired clean-twin protocol is defined in Section 3; the target estimand τ_{int}, the paired estimator c_{int}, the DiD estimator $c_{\mathrm{int}}^{\mathrm{DiD}}$, and the reported-cost quantity c_{def} are given in closed form. All models are public (Pythia-410M, Qwen2.5-7B-Instruct, Llama-3.2-3B-Instruct, Llama-2-7B-Chat, Mistral-7B, ResNet-18, PreAct-ResNet-18, VGG19-BN, and CGD's CLIP ViT-B/32) and all datasets, benchmark suites and toolkits are public (the Pile subset, Alpaca-52k, MMLU, CIFAR-10, CIFAR-100, TOFU, BackdoorBench, BackdoorBox). The appendices report full defense grids, hyperparameters, per-defense summaries, success counts and pooled intervals; together with the supplementary code and raw per-installation data they provide the complete poison-generation, twin-construction, training, defense and evaluation procedures.

REFERENCES

Gorka Abad, Marina Krček, Stefanos Koffas, Behrad Tajalli, Marco Arazzi, Roberto Riaño, Xiaoyun Xu, Zhuoran Liu, Antonino Nocera, and Stjepan Picek. SoK: The last line of defense: On backdoor defense evaluation. *arXiv preprint arXiv:2511.13143*, 2025.

Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in CNNs by training set corruption without label poisoning. In *IEEE International Conference on Image Processing (ICIP)*, 2019. arXiv:1902.11237.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, et al. Measuring what matters: Construct validity in large language model benchmarks. *arXiv preprint arXiv:2511.04703*, 2025.

Lisa Bouger, Théo Lasnier, Philippe Loubet Moundi, Yannick Teglia, and Djamé Seddah. Backdoor unlearning generalization: A path toward the removal of unknown triggers in LLMs. *arXiv preprint arXiv:2606.03785*, 2026.

Xavier Bouthillier et al. Accounting for variance in machine learning benchmarks, 2021.

Hoyeon Chang et al. How do large language models acquire factual knowledge during pretraining? *arXiv preprint arXiv:2406.11813*, 2024.

Weixin Chen, Baoyuan Wu, and Haoqian Wang. Effective backdoor defense by exploiting sensitivity of poisoned samples. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Sungjun Cho, Dasol Hwang, Frederic Sala, Sangheum Hwang, Kyunghyun Cho, and Sungmin Cha. Reference-specific unlearning metrics can hide the truth: A reality check. *arXiv preprint arXiv:2510.12981*, 2025.

Dami Choi, Yonadav Shavit, and David Duvenaud. Tools for verifying neural models' training data. *arXiv preprint arXiv:2307.00682*, 2023.

Lee J. Cronbach and Lita Furby. How we should measure "change"—or should we? *Psychological Bulletin*, 74(1):68–80, 1970.

Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. LIRA: Learnable, imperceptible and robust backdoor attacks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2410.07163.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *ACM Conference on Recommender Systems (RecSys)*, 2019. arXiv:1907.06902.

Polina Gordienko, Georg Schollmeyer, Frauke Kreuter, and Christoph Jansen. How hard is it to rig a benchmark? a social choice analysis of leaderboard robustness. *arXiv preprint arXiv:2605.23628*, 2026.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

Jackson Harmon, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. Mapping post-training forgetting in language models at scale. *arXiv preprint arXiv:2510.17776*, 2025.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Abdullah Ahmad Khan, Hamid Laga, and Ferdous Sohel. Metric unreliability in multimodal machine unlearning: A systematic analysis and principled unified score. *arXiv preprint arXiv:2605.02206*, 2026.

Yongwoo Kim, Sungmin Cha, and Donghyun Kim. Are we truly forgetting? a critical re-examination of machine unlearning evaluation protocols. *Engineering Applications of Artificial Intelligence*, 167, 2026. arXiv:2503.06991.

Martin Klein, Tommy Wright, and Jerzy Wieczorek. Confidence intervals for ranks, 2018.

Jianwei Li and Jung-Eun Kim. Purifying generative LLMs from backdoors without prior knowledge or clean reference. *arXiv preprint arXiv:2603.13461*, 2026.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a. arXiv:2110.11571.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations (ICLR)*, 2021b.

Yige Li, Xixiang Lyu, Xingjun Ma, Nodens Koren, Lingjuan Lyu, Bo Li, and Yu-Gang Jiang. Reconstructive neuron pruning for backdoor defense. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2305.14876.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021c. arXiv:2012.03816.

Weilin Lin, Li Liu, Shaokui Wei, Jianze Li, and Hui Xiong. Unveiling and mitigating backdoor vulnerabilities based on unlearning weight changes and backdoor activeness. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.20291.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *RAID*, 2018a. arXiv:1805.12185.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Network and Distributed System Security Symposium (NDSS)*, 2018b.

Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *European Conference on Computer Vision (ECCV)*, 2020. arXiv:2007.02343.

Frederic M. Lord. A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68 (5):304–305, 1967.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs, 2024.

Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid Hasan, and Elita Lobo. Alternate preference optimization for unlearning factual knowledge in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pp. 3732–3752, 2025.

Nay Myat Min, Long H. Pham, and Jun Sun. Unified neural backdoor removal with only few clean samples through unlearning and relearning. *IEEE Transactions on Information Forensics and Security*, 2024. arXiv:2405.14781.

Rui Min, Zeyu Qin, Li Shen, and Minhao Cheng. Towards stable backdoor purification through feature shift tuning. In *NeurIPS*, 2023. arXiv:2310.01875.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision (ECCV)*, 2020. arXiv:2003.08505.

Milad Nasr, Nicholas Carlini, Chawin Sitawarin, Sander V. Schulhoff, Jamie Hayes, Michael Ilie, Juliette Pluto, Shuang Song, Harsh Chaudhari, Ilia Shumailov, Abhradeep Thakurta, Kai Yuanqing Xiao, Andreas Terzis, and Florian Tramèr. The attacker moves second: Stronger adaptive attacks bypass defenses against LLM jailbreaks and prompt injections. *arXiv preprint arXiv:2510.09023*, 2025.

Bitya Neuhof and Yuval Benjamini. Rank intervals for leaderboards: A hierarchical framework for model evaluation, 2026.

Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2010.08138.

Anh Nguyen and Anh Tran. WaNet – imperceptible warping-based backdoor attack. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2102.10369.

Stanisław Pawlak, Jan Dubiński, Daniel Marczak, and Bartłomiej Twardowski. Backdoor vectors: A task arithmetic view on backdoor attacks and defenses. *arXiv preprint arXiv:2510.08016*, 2025.

Jie Peng, Hongwei Yang, Jing Zhao, Hengji Dong, Hui He, Weizhe Zhang, and Haoyu He. Circumventing backdoor space via weight symmetry. *arXiv preprint arXiv:2506.07467*, 2025.

Qi Qian, Chengsong Huang, Jingwen Xu, Changze Lv, Muling Wu, Wenhao Liu, Xiaohua Wang, and Xiaoqing Zheng. Benchmark^2: Systematic evaluation of LLM benchmarks. *arXiv preprint arXiv:2601.03986*, 2026.

Mark Rofin, Aditya Varre, and Nicolas Flammarion. (how) learning rates regulate catastrophic overtraining. In *Conference on Language Modeling (COLM)*, 2026. arXiv:2604.13627.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2304.15004.

Rylan Schaeffer, Joshua Kazdan, Baber Abbasi, Ken Ziyu Liu, Brando Miranda, Ahmed Ahmed, Fazl Barez, Abhay Puri, Stella Biderman, Niloofar Mireshghallah, and Sanmi Koyejo. Quantifying the effect of test set contamination on generative evaluations. *arXiv preprint arXiv:2601.04301*, 2026.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, et al. MUSE: Machine unlearning six-way evaluation for language models, 2024.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025.

Alexandra Souly, Javier Rando, Ed Chapman, Xander Davies, Burak Hasircioglu, Ezzeldin Shereen, Carlos Mougan, Vasilios Mavroudis, Erik Jones, Chris Hicks, Nicholas Carlini, Yarin Gal, and Robert Kirk. Poisoning attacks on LLMs require a near-constant number of poison samples. *arXiv preprint arXiv:2510.07192*, 2025.

Bo Su, Ankit Shah, and Thai Le. PreUnlearn: Auditing collateral knowledge damage before large language model unlearning. *arXiv preprint arXiv:2606.18473*, 2026.

Jingtong Su, Jianyu Zhang, Karen Ullrich, Léon Bottou, and Mark Ibrahim. A single character can make or break your LLM evals. *arXiv preprint arXiv:2510.05152*, 2025.

Luze Sun, Anshuman Suri, Harsh Chaudhari, Cristina Nita-Rotaru, and Alina Oprea. PoisonForge: Task-level targeted poisoning benchmark for instruction-tuned LLMs. *arXiv preprint arXiv:2605.23168*, 2026.

Yao Tong, Weijun Li, Xuanli He, Haolan Zhan, and Qiongkai Xu. Cut the deadwood out: Backdoor purification via guided module substitution. *arXiv preprint arXiv:2412.20476*, 2025.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, 2019.

Tong Wang, Yuan Yao, Feng Xu, Shengwei An, Hanghang Tong, and Ting Wang. An invisible black-box backdoor attack through frequency domain. In *European Conference on Computer Vision (ECCV)*, 2022a.

Zhenting Wang, Juan Zhai, and Shiqing Ma. BppAttack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b. arXiv:2205.13383.

Shaokui Wei, Mingda Zhang, Hongyuan Zha, and Baoyuan Wu. Shared adversarial unlearning: Backdoor mitigation by unlearning shared adversarial examples. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2307.10562.

Shaokui Wei, Hongyuan Zha, and Baoyuan Wu. Mitigating backdoor attack by injecting proactive defensive backdoor. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.16112.

Shaokui Wei, Jiayin Liu, and Hongyuan Zha. Backdoor mitigation by distance-driven detoxification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. arXiv:2411.09585.

Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, Mingli Zhu, Ruotong Wang, Li Liu, and Chao Shen. BackdoorBench: A comprehensive benchmark and analysis of backdoor learning. *International Journal of Computer Vision*, 2025. arXiv:2407.19845.

Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Binyan Xu, Fan Yang, Xilin Dai, Di Tang, and Kehuan Zhang. CLIP-guided backdoor defense through entropy-based poisoned dataset separation. In *ACM International Conference on Multimedia (ACM MM)*, 2025. arXiv:2507.05113.

Jiatong Yu, Yinghui He, Anirudh Goyal, and Sanjeev Arora. On the impossibility of retrain equivalence in machine unlearning, 2025.

Yi Zeng, Won Park, Z. Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. arXiv:2104.03413.

Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2110.03735.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning, 2024.

Zhiyuan Zhang et al. Fine-mixing: Mitigating backdoors in fine-tuned language models. *Findings of EMNLP*, 2022. arXiv:2210.09545.

Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:2005.00060.

Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Data-free backdoor removal based on channel Lipschitzness. In *European Conference on Computer Vision (ECCV)*, 2022. arXiv:2208.03111.

Jun Yen Zhu, Paul Michel, Huan Yao, and Graham Neubig. Rethinking backdoor detection evaluation for language models. In *EMNLP (Main)*, 2025. arXiv:2409.00399.

Mingli Zhu, Shaokui Wei, Li Shen, Yanbo Fan, and Baoyuan Wu. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *ICCV*, 2023a. arXiv:2304.11823.

Mingli Zhu, Shaokui Wei, Hongyuan Zha, and Baoyuan Wu. Neural polarizer: A lightweight and effective backdoor defense via purifying poisoned features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b. arXiv:2306.16697.

S1 WHAT THE AUDITED PAPERS REPORT

We read the six papers that define the defenses and the rating this audit uses, and recorded three things from each: the clean-accuracy cost it reports for its own headline defense, whether it ever applies that defense to a never-poisoned model, and whether it states or varies the defender's clean-data budget. Table 5 is the result. Two patterns hold across all six. None runs the defended clean arm: the closest any comes is quoting a clean model's accuracy *level* as a reference point (Wang et al., 2019, 77.2%; Wu et al., 2025, Table 3), never its cost under the defense. And where the budget is varied it is varied as a robustness check on the defense, not as a confounder in the cost column—Zhu et al. (2023a) even report the rating itself at three budgets without reading the movement as a property of the budget.

The reported costs span two regimes. Each paper's own headline defense is cheap—0.2 to 3.6 points, and Zhu et al. (2023a) report a net *gain* of 0.86—which is the band drawn in Figure 4(b). The same tables carry other defenses that are not: the widest entry we verified is ABL at 47.5 points in Zhu et al. (2023a)'s Table 1, independently reproduced in Wu et al. (2025)'s Table 14, with AC at 46.1 and Neural Cleanse at 34.6 in the same table. Without the defended clean arm, a column that contains both regimes cannot isolate removal-attributable cost from intervention cost.

Table 5: Audit of the six papers whose defenses and rating this work uses. "Headline cost" is the clean-accuracy drop the paper reports for its own primary defense, in accuracy points, at the location given. "Clean arm" asks whether the paper applies its defense to a never-poisoned model and reports the resulting cost. †the last row is a benchmark suite, not a single proposed defense; its column gives the per-defense mean range with the largest single entry.

Audited entry	defense	reported cost (pt)	clean arm?	clean-data budget
Liu et al. (2018a)	fine-pruning	0.2 (Tab. 1)	no	stated only, size not given
Li et al. (2021b)	NAD	2.66 (Tab. 1)	no	5%; swept 0–20% (Fig. 2)
Wu & Wang (2021)	ANP	1.47 (Tab. 1)	no	1%; swept 0.1–10% (App.)
Wang et al. (2019)	unlearning	3.6 max (Tab. IV)	no	10%, not swept
Zhu et al. (2023a)	FT-SAM	−0.86 (gain; Tab. 1)	no	5%; swept 1–10% (Tab. 3)
Wu et al. (2025)†	nine-defense suite	0.05–24.4 mean; 47.5 max (Tab. 14)	no	5% fixed, not swept

Table 6: What the nearest prior work reports, from each paper's own text. ✓: reported; ○: partial; —: not reported. The defended clean arm is not our novelty: three of the six run one and the systematization recommends it (its R7). What none of them does is the rest of the row: define the target with identification conditions, net the pre-defense gap, condition on removal success, replicate at the installation level with intervals, carry the correction into the field's ranking rule, and test a matched-twin-free substitute. Locations of the supporting sentences are in the supplementary audit notes.

Work	defended clean arm	interaction estimand + identification	nets pre-defense gap	success-conditioned	independent installations	ranking consequence	surrogate for the twin
Abad et al. (2025) (SoK)	✓[a]	—	—	—	○[b]	—	—
Peng et al. (2025)	✓[c]	—	—	○[d]	—	—	—
Tong et al. (2025)	✓[c]	○[e]	—	—	○[f]	—	—
Bouger et al. (2026)	—[g]	—	—	○[d]	—	—	—
Li & Kim (2026)	—	—	—	—	—	○[h]	○[i]
Cho et al. (2025) (unlearning)	○[j]	—	—	—	✓	✓[k]	—
TARE (this paper)	✓	✓	✓	✓	✓	✓[l]	✓

[a] poison-rate-0 arm with absolute defended clean accuracy, no delta; [b] ten repeated trials in one configuration, main results single-run; [c] appendix only, raw accuracy, never subtracted from the poisoned-arm cost; [d] a success threshold or final attack success is shown, cost is reported unconditionally; [e] Δ_{acc} against the victim model as a search objective; [f] three seeds averaged, no intervals, replicate unit unstated; [g] its control is the backdoored model continually pretrained on clean data, and capability is compared to the base level; [h] a prose re-run of one baseline at standard learning rates; [i] reference-free by design, through paired variants derived from the suspicious model, not an independent clean model; [j] a retain-only gold model as the forgetting reference, utility reported as a level; [k] on the forgetting axis; [l] in the audited CIFAR-10/BadNets suite.

S2 NUISANCE SENSITIVITIES

The language-model setting. A lexical backdoor (a rare-token trigger mapping to a fixed payload) is installed into Qwen2.5-7B-Instruct and Llama-3.2-3B-Instruct by continued training on poisoned data, with a clean twin per seed; Q is MMLU accuracy, and the two defenses are budget points, not deployable operating points (Appendix S8, Table 7). θ_{base} in Eq. 2 is the pre-defense poisoned model in both modalities, the vision convention and the conservative choice.

Table 7: Table 25 with the language-model row. Qwen2.5-7B: eight installations, two defenses (continued fine-tuning and fine-pruning at budget points that are not deployable operating points, Appendix S8); $^{\diamond}$its pooled interaction resolves above zero only through an $r = -0.90$ anti-correlation of the two defenses over 6 installations; under independence it is $[-0.31, +3.05]$. Llama (one installation, three defense seeds): Table 8.

Setting (unit)	A3	c_{def}	tare c_{clean}	g	DiD	paired
BackdoorBench, 11 seeds*	$74/77^{\S}$	$+2.34$ [+2.10,+2.57]	$+1.98$ [+1.50,+2.46]	$+2.06$	$+0.36$ [−0.26,+0.98]	$+2.42$ [+1.81,+3.04]
our rig, 8 installations	$37/48$	$+2.51$ [+2.25,+2.77]	$+2.49$ [+2.03,+2.96]	$+0.41$	$+0.02$ [−0.31,+0.34]	$+0.43$ [+0.30,+0.55]
framework recovery schedule¶	$42/48$	$+2.20$ [+2.04,+2.36]	$+2.24$ [+1.84,+2.63]	$+0.41$	-0.04 [−0.30,+0.23]	$+0.38$ [+0.22,+0.53]
Qwen2.5-7B (8 installations)	$14/16$	$+11.46$ [+8.94,+13.97]	$+10.08$ [+7.61,+12.56]	$+0.04$	$+1.37$ [+0.52,+2.23]$^{\diamond}$	$+1.42$ [+0.69,+2.14]

Table 8 varies one analysis decision at a time and reports the same two quantities as Table 25; the admissibility axes are separated into Appendix S3.

Table 8: **At a fixed attack the cost gap is stable across the replicated admissible alternatives and the one coordinate that nearly closes it belongs to the evaluator; across attacks it is not even sign-stable.** Under the shared convention of Table 29 the gap—c_{def} minus the DiD removal term, which is the clean arm—is $+1.98$ [+1.50,+2.46] on BadNets and -0.55 [−0.90,−0.20] on WaNet, so it is a property of the (attack, defense-budget) pair and not a constant (Section 4.6). The WaNet row below pools eleven installation seeds; the -0.05 an earlier version quoted as "the" gap was seed 0 alone. Rows vary one analysis or setting decision each and report the cost decomposition, not the induced ordering, which is treated separately in Section 4.5. Each row reports c_{def} and the removal estimate used (Appendix S5: not identified); sources in Appendix S5 and Appendices S16–S30. *this row's cost column is c^0_{def} of Section 3.3, the clean arm's own drop, not c_{def}. †WaNet: eleven installation seeds under the same frozen protocol. Three $(3, 8, 10)$ fail A0—the attack installs at only 0.37–0.75 success against 0.94–0.99 on the other eight—and the success-conditioned common support is five seeds over six candidate defenses. Its poisoned model starts 1.43 $[+0.48, +2.39]$ pt below its twin, the framework's 100 recovery epochs out-train its attack script on $5/5$ common-support seeds for every *candidate* defense but ANP (over all eight, ABL and Neural Cleanse break it), and $5/6$ candidates report a *negative* cost, each interval excluding zero, whose clean arm carries -20% to 67%—I-BAU is the one whose clean arm carries the *opposite* sign to its reported cost. The arrow in the WaNet row is the convention of Table 29 (seven defenses excluding ABL, A3-failed cells dropped, eleven seeds on both grids); under the matched candidate set, success-conditioned, it is $+2.44$ [+2.18,+2.70] $\rightarrow -1.32$ [−2.18,−0.46], so the sign change survives either. ‡TOFU's figure is the retrain-gold contrast of Appendix S23, not the four-cell DiD. ¶language model then vision, the correction of Appendix S30; the pre-fix values appear only there and here.

sensitivity	cost column	removal estimate (form)	effect on conclusion
estimator: DiD → paired	$+2.51$	$+0.02 \rightarrow +0.43$	paired resolves, DiD does not
clean-data budget: 1% → 100%	$+3.80 \rightarrow -0.05$	$+0.01 \rightarrow -0.33$ (DiD)	removal within ± 0.4
same sweep, no-backdoor arm*	$3.79 \rightarrow 0.28$	$\equiv 0$ by construction	tax alone spans the band
attack: BadNets → blended (8 seeds)	per defense	per defense	share 43–138%; 1 of three pairs reverse, zero survive
attack: BadNets → WaNet (11 seeds)†	$+2.34 \nrightarrow -1.00$ [−1.53,−0.47]	-0.45 [−1.03,+0.13] (DiD)	cost changes *sign*; 2 of 20 reverse, zero survive
headroom route: noise → data shortage	—	$-1.14 \rightarrow +0.41$ at g=2.06	not identified; we apply none
defense family: 4 always-clearing configs	$+1.12$–$+4.28$	-0.14–$+0.12$ (DiD)	unchanged; none resolves
model/modality: four settings	11.5/9.0/2.5/publ.	$+1.37/-1.73/+.02$; $+.82$‡	removal small and signed
recovery schedule: 10 → 100 epochs	$+2.51 \rightarrow +2.20$	$+0.02 \rightarrow -0.04$ (DiD)	unchanged; A3 $37 \rightarrow 42$ of 48
defender draw: shared → independent¶	$8.8 \rightarrow 11.5$; $3.4 \rightarrow 2.8$	$-.24 \rightarrow +1.37$; $+1.05 \rightarrow +.02$	both signs change
twin → surrogate: 4 installations	—	-0.02 paired; -0.03 DiD	unchanged within recipe
prune → output-masked prune	$+2.81 \rightarrow +2.51$	$+0.04 \rightarrow +0.02$ (DiD)	unchanged
ANP → reference mask and threshold	$+8.3 \rightarrow +2.0$	$+0.25$ [−0.71,+1.22]	drops on A3, clears $3/14$
NC → pixel-space reversal	$+4.24 \rightarrow +0.78$	-0.14 [−0.97,+0.70]	becomes a candidate, $14/14$
headroom arm at a matched deficit	$\equiv 0$ by construction	-0.28 [−0.63,+0.08]	triple difference $+0.32$

S3 ADMISSIBILITY: SUCCESS RATES, THRESHOLD, AND COMMON SUPPORT

In full. On our own suite the ordering consequence is arithmetic, so we do not report it as a result. The removal term varies by 0.11 points across our five candidates against 1.38 for the clean arm, so an ordering by reported cost almost has to be one by intervention, and it is exactly that: the two agree on all five, $10/10$ pairs separable (Figure 7). The separable reversals that count are on BackdoorBench (Section 4.6).

Eq. 1 conditions on successful removal, so the protocol has two outputs and we report both. Table 9 gives the success rate of each vision defense over the fourteen installations at the operational threshold $\alpha = 0.05$; the threshold is ours, chosen before the ranking analysis to sit an order of magnitude below the 96–97% attack success the undefended models carry, not inherited from a benchmark. Five methods pass the gate; four of the five clear the backdoor in every one of the fourteen installations and NAD misses one, so the success-conditioned common support is thirteen of fourteen.

The clean arm is not sign-stable across attacks, and that is the strongest form of our claim. Running the same eight defenses over eleven seeds against WaNet instead of BadNets, with the same unconditional twin-arm convention (all eleven seeds, no success criterion), inverts five of the eight in the mean and three with intervals excluding zero: fine-tuning -1.99 $[-2.30,-1.68]$, fine-pruning -1.68 $[-2.12,-1.24]$ and NAD -1.85 $[-2.14,-1.56]$, against $+0.13$, $+1.05$ and $+0.33$ on BadNets. ANP and ABL (both $A2$ at the policy level only) keep a sign whose interval excludes zero ($+1.78$ $[+0.52,+3.04]$, $+27.77$ $[+21.27,+34.26]$) and the remaining three straddle zero. Pooled over the six candidates on that same unconditional convention the clean arm is -0.61 $[-0.93,-0.29]$, an interval that excludes zero; success-conditioned on the five-seed common support it is -0.69 $[-1.46,+0.09]$. **This is a counterexample to the unconditional-universal form of our own headline**—"every defense charges a never-poisoned model" holds on BadNets, not on WaNet—and we report it in that form rather than restricting the claim quietly. It is also the sharper version of the paper's thesis: a quantity that flips sign with the victim's training schedule cannot be a fixed overhead that the audited protocols implicitly netted out, and it cannot be recovered from the poisoned arm alone. The mechanism is visible in the same grid: BackdoorBench's 100 recovery epochs out-train WaNet's attack script on $5/5$ common-support seeds for every candidate defense but ANP, so the defended poisoned arm ends *above* where it started. The mechanism is in the starting point, and it is visible in the same cells: WaNet's rig leaves *both* arms about two points lower than BadNets' does (twin 0.9150 against 0.9371, poisoned 0.9005 against 0.9164), so the framework's 100 recovery epochs on a clean subset are net *training* on the WaNet grid and net damage on the BadNets grid. The same fine-tuning that costs the BadNets twin 0.13 points raises the WaNet twin by 1.99, and raises its poisoned arm by 3.10. Only the two defenses that damage a model regardless of where it starts—ANP, which prunes, and ABL, which unlearns—keep a sign whose interval excludes zero on both attacks. **So the clean arm is the difference between where the attack's training run left the victim and the fixed point the defense's own recovery run converges to.** That fixed point is visible in the four cells of Appendix S21: the undefended twin moves 3.10 points across them while the defended twin moves at most 0.66 (fine-tuning 0.30, NAD 0.29, fine-pruning 0.66)—a level set by the defense, not by where the victim started. And the poisoned arm's excess over its twin does not move with the schedule, the start gap returned rather than removal identified (Section 4.2, Appendix S5): on WaNet, DiD is -1.11 $[-1.64,-0.58]$ under the shipped schedule and -0.94 $[-1.38,-0.49]$ under the swapped one for fine-tuning, -1.09 $[-1.57,-0.62]$ and -1.00 $[-1.46,-0.55]$ for NAD, every pair overlapping; on the un-annealed BadNets grid only $15/33$ poisoned-arm cells clear A3 and its residue, -0.74 $[-1.62,+0.13]$ against $+0.46$ $[+0.27,+0.65]$, is not read. What the schedule moves is the tare, which is why it cannot be a constant anyone subtracted after the fact, and it is not recoverable from the poisoned arm alone—which is the whole point of measuring the twin.

A0 is not a formality: on one attack it excludes 27% of seeds. On our CIFAR grid and on BackdoorBench's BadNets, every seed installs, so A0 never binds and a reader could reasonably ask what it is for. On WaNet it binds hard. Running the same frozen protocol over eleven installation seeds, three (seeds 3, 8, 10) reach a pre-defense attack success of only 0.641, 0.369 and 0.751, against 0.938–0.988 on the other eight; their clean accuracy is normal (0.895–0.907), so nothing but the trigger check distinguishes them. Conditioning on removal success is meaningless on a model that was never backdoored to begin with, so A0 removes them, and the success-conditioned common support falls to five seeds. **The consequence is about the literature, not about us**: a cost or ranking

read from a *single* installation of this attack has a better than one in four chance of being read off a model the attack did not take on, and the audited protocols we survey report neither a per-seed installation rate nor an installation gate. Our own earlier single-seed WaNet reading happens to have landed on an installing seed—seed 0, which the eleven-seed grid reproduces exactly—but that was luck, and the eleven-seed interval, not the single reading, is what we now report.

Table 9: Installation-level success rates and residual attack success for all nine defenses, over fourteen independent CIFAR-10 installations with the corrected, independent defender draw, fine-pruning that silences the pruned channel at the layer output, ANP at its reference mask semantics under its own accuracy-guarded threshold, and Neural Cleanse reverse-engineering its trigger in pixel space. Five methods pass the candidate gate (method-level mean residual $\leq \alpha = 0.05$); four of them clear the backdoor in every installation, and every twin is admissible throughout. Regenerated from `data/cifar_cells.csv` by `analysis/build_a3_table.py`, which also re-derives the candidate set and the common support and asserts them against the flags shipped in the cell file. [§]NAD is a candidate on its method-level mean residual (0.022) but misses installation 3 (0.138), which is what drops the ranking's common support from fourteen to thirteen; this is the corrected teacher of Appendix S30, not the lr 0.01 teacher our first implementation used. [¶]ANP under a two-point clean-accuracy budget prunes too little to clear this attack: it costs only $+2.0$ points but leaves a mean residual attack success of 0.51, so it is excluded on removal success, not on twin admissibility. Pruning it to a fixed 20% instead clears 5 of 8 at $+7.7$ points and destroys 2 of 8 (Appendix S24).

Defense	successes (A3)	defended twin clears	mean residual A	max residual A
Neural Cleanse	14/14	14/14	0.008	0.012
fine-prune 40%	14/14	14/14	0.014	0.019
fine-prune 30%	14/14	14/14	0.016	0.045
fine-tune lr 0.05	14/14	14/14	0.018	0.021
NAD[§]	13/14	14/14	0.022	0.138
fine-prune 20%	10/14	14/14	0.062	0.318
ANP[¶]	3/14	14/14	0.509	0.965
fine-tune lr 0.02	1/14	14/14	0.613	0.914
fine-tune lr 0.01	0/14	14/14	0.856	0.940

What success-conditioning moves, and what it does not. A3 filters on a post-treatment variable measured on the poisoned arm only, so it can in principle select on the outcome. Empirically it moves the reported cost and leaves the residue alone: on the CIFAR headline grid the conditional reported cost is $+2.51$ $[+2.25, +2.77]$ against $+2.11$ $[+1.99, +2.22]$ unconditional, while the interaction is $+0.02$ $[-0.31, +0.34]$ against $+0.04$ $[-0.24, +0.32]$; on Qwen the two interactions agree to two decimals. Conditioning also makes the per-installation average run over a data-dependent set of configurations—installations retain four to six of the six—so part of the reported cost's spread is which defenses happened to clear in that seed. Restricting to the four configurations that clear in every installation removes that source and gives $+2.73$ $[+2.61, +2.85]$ reported and -0.00 $[-0.37, +0.36]$ removal, which is the balanced reading of the same grid.

The rating's clip, and the ordering without it. The $\max(0, \cdot)$ clip in Eq. 7 zeroes any negative capability drop, and under the DiD accounting it binds in 37 of the 65 candidate cells: NAD in 10 of 13, fine-tuning at lr 0.05 in 9 of 13, fine-pruning at 40% in 7 of 13, fine-pruning at 30% in 7 of 13, Neural Cleanse in 4 of 13. That is why the five corrected ratings lie within 0.003. Removing the clip changes the corrected ordering—τ_a(published, removal) falls from $+0.800$ to $+0.600$ and the gap widens from $+0.200$ to $+0.400$, and a second pair inverts with it (fp 40 4→3, fp 30 3→5, ft .05 5→4)—so the corrected *order* is itself partly an artifact of the rating's shape. The published-versus-intervention agreement is $\tau_a = +1.000$ with and without the clip, which is the part of the ordering result that does not depend on it.

The utility floor on A3. A3 asks that the residual attack success fall below α, which a defense can satisfy by destroying the model: a network at chance accuracy answers the trigger no more often than anything else. We therefore also require the defended arm to retain capability—clean accuracy within 20 points of the pre-defense model on vision, MMLU above chance plus ten points on the language models. No cell in the paper's own grids is excluded by this floor except 1 run in the

BadNets dose grid of Appendix S51 (SAU at 5%, seed 2) and 7 ABL runs in its Blended/LF grid, none in a printed column; in the leaderboard grids of Appendix S52 it keeps out of A3 the poisoned-arm runs whose residual attack success was below α only because the defense had destroyed the model: 2 runs of the seven benchmark defenses other than ABL (Trojan/SAU seed 3, SIG/SAU seed 2), 48 of RNP's, and, from the A3 counts of the failure records, 14 of ABL's and 18 of D-BR's. It is not idle: an earlier implementation of ANP produced cells at 10% clean accuracy that A3 recorded as successful removals (Appendix S30), and a success criterion that cannot tell removal from destruction is a gap in the estimand, not in the implementation.

The gate does not make the result. Sweeping α changes the candidate set and its common support (Table 10). At the operational $\alpha = 0.05$ the candidate set is five methods on a common support of thirteen and the gap is $+0.200$, for which we report two intervals rather than one, because the two standard bootstrap constructions on the same $5{,}000$ installation resamples disagree about zero: percentile $[+0.200, +0.600]$, basic (reverse-percentile) $[-0.200, +0.200]$. The disagreement is a property of the statistic, not a choice we make: Kendall τ_a on five methods moves in steps of 0.200, the observed gap is one step, and it sits at the 0.9th percentile of its own resample distribution (median $+0.400$, mean $+0.381$, so the bootstrap bias $+0.181$ is as large as the estimate), which pins the percentile interval's lower endpoint on the point estimate itself—the diagnostic of a failed percentile interval, not evidence of an effect bounded away from zero. What the resamples support is the sign and not the magnitude: none returns a negative gap, 0.9% return exactly zero ($P^*(\text{gap} \leq 0) = 0.0086$), and dropping any single installation leaves the gap in $[+0.200, +0.400]$. We therefore report the gap as positive with an unresolved magnitude, and make no claim that its interval excludes zero. Tightening to $\alpha = 0.02$ drops NAD and leaves four methods on all fourteen with a wider gap whose intervals touch zero ($+0.333$, percentile $[+0.000, +0.667]$); loosening to $\alpha = 0.10$ admits fine-pruning at 20%, drops the common support to nine (installations 3, 21, 23, 27 and 29 drop out: NAD leaves 0.138 on one and fine-pruning at 20% fails on four more), and widens the gap to $+0.667$ (percentile $[+0.400, +1.200]$). The gap is positive at every threshold; its percentile interval excludes zero at three of the four rows and its basic interval at two, and the one row where the two constructions disagree is the operational $\alpha = 0.05$ (Table 10 prints both), so no threshold row supports an unqualified "excludes zero". At every threshold the published ordering coincides with the intervention-component ordering ($\tau_a = +1.000$).

Table 10: Candidate-gate sensitivity. Sweeping the method-level mean-residual-ASR threshold α over the fourteen installations changes the candidate set, and with it the common support the ranking is computed on; it does not change the installation-level A3 event. Each row is the primary analysis of Appendix S18 run at that α: Kendall τ_a of the published ordering against each component on that row's own common support, with an installation bootstrap on the gap. Both bootstrap constructions are printed, and the common support each row runs on, because the two constructions disagree about zero on the operational row. The rows are four distinct analyses: $\alpha = 0.02$ excludes NAD on its mean residual and so runs on four methods over all fourteen installations, $\alpha = 0.05$ (the operational threshold, and the one every other number in the paper uses) admits NAD and therefore runs on five methods over the thirteen installations NAD clears, and the two loosest rows readmit fine-pruning at 20%. The all-fourteen-installation variant of the $\alpha = 0.05$ row is the last line of Table 13.

α	methods	common support	τ_a(published, tax)	τ_a(published, removal)	gap	percentile 95%	basic 95%
0.02	4	14	+1.000	+0.667	+0.333	[+0.000,+0.667]	[+0.000,+0.667]
0.05	5	13	+1.000	+0.800	+0.200	[+0.200,+0.600]	[−0.200,+0.200]
0.10	6	9	+1.000	+0.333	+0.667	[+0.400,+1.200]	[+0.133,+0.933]
0.25	6	9	+1.000	+0.333	+0.667	[+0.400,+1.200]	[+0.133,+0.933]

Common support is the primary convention. Four of the five candidate methods clear the backdoor in every installation here and NAD misses one, so the common support is thirteen of the fourteen; the convention matters both for that installation and because a looser gate readmits methods that clear far less often. For a candidate set $\mathcal{D}$ define $\Omega_{\text{CS}}(\mathcal{D}) = \{\omega : A(\mathcal{D}_d\theta^1(\omega)) \leq \alpha \ \forall d \in \mathcal{D}\}$, the installations on which every method in $\mathcal{D}$ clears the backdoor. Appendix S18 restricts primary cross-method inference to Ω_{CS}—thirteen of the fourteen installations here, the one exception being the installation NAD does not clear. This is a *ranking-specific* extension of Eq. 1, not the single-method estimand itself: Algorithm 1 handles one defense at a time and does not define the intersection. It places every method on the same runs, at the price of depending on $\mathcal{D}$. Pairwise separability there follows Section 3.4: installations are resampled as clusters with Ω_{CS} fixed, each

pair's two-sided bootstrap p-value for the sign of its rating difference is corrected by Benjamini–Hochberg at $q = 0.05$ within each accounting's own set of pairwise comparisons, and rank intervals are a complementary display. As a null-centered check, an exact enumeration of the paired sign-flip test under the sign-symmetry null on the thirteen per-installation rating differences ($8{,}192$ sign assignments per pair) returns $10/10$ published-accounting pairs and $3/10$ corrected pairs after the same Benjamini–Hochberg step, matching the bootstrap exactly. A4 is a property of the *undefended* twin, $A(\theta^0) \leq \alpha$, and holds for $14/14$ vision twins (attack success $\leq 0.8\%$ before any defense) and $8/8$ Qwen twins (0); the shipped cells now carry the undefended twin's attack success as its own column, since the flag we had named A4 in them was computed from the defended twin—the quantity Table 9 reports—and both hold in $14/14$, so no number moves. The one A4 failure, the blended-trigger twin at the benchmark's 0.0002 rate, is not identified as a control; blended's row in Table 8 uses the one-image twin that passes A4 (Appendix S20). Whether a twin drawn at a fixed poisoning rate is clean is a property of the attack, not of the rate: evaluating three models on one attack's backdoor test set gives the twin's own score, the score of a same-recipe model that never saw that trigger, and the poisoned model's, and the first minus the second is what those ten images installed—0.001 for BadNets and 0.004 for WaNet, both on seed 0, against 0.108 for blended averaged over three (`supplement/data/twin_admissibility_by_attack.csv`); over the whole eleven-seed WaNet grid the undefended twin's attack success runs 0.0073–0.0178 (mean 0.0135) and A4 holds $11/11$. Table 9 additionally reports, per defense, whether the *defended* twin still clears the threshold, which is what a reader needs to know before treating that defense's clean arm as a control; under the corrected implementations it does so in every cell.

When each choice was fixed. Table 11 records the status of every analysis choice in the ranking chain, separating decisions fixed before the relevant data were inspected from those adopted during internal review and reported with their alternatives.

Table 11: Analysis choices behind the ranking result and when each was fixed. "Internal review" denotes revisions made after a first all-nine, method-level analysis had been written up and before submission; every such revision is reported beside the alternative it replaced.

choice	status	sensitivity reported
installation-level A3 threshold $\alpha = 0.05$	fixed before the ranking analysis	Table 10
DER as the field's accounting rule	fixed (Zhu et al., 2023a)	—
equal weighting of configurations within an installation	fixed from the first aggregate	per-defense rows, complete case
independent installation as the replicate unit	adopted during internal review; the eight-installation Qwen rig was run after it	defense-seed rows marked
method-level mean-ASR candidate gate	adopted during internal review after the all-nine ranking	all-nine result retained (Table 31)
common-support primary convention	adopted during internal review after the fourteen-installation grid	two looser conventions (Table 13)
installation-cluster bootstrap and BH families	fixed before the pairwise analysis	—
exact sign-flip enumeration	post hoc sensitivity	reported above
ANP kept in the pooled TARE-Z summary	fixed	excluding-ANP rows (Table 35)
ANP's pruning threshold: the method's accuracy budget	adopted when the implementation was corrected	fixed-fraction rows (Appendix S24)

Table 13 reports that ranking beside the two looser conventions. The published-versus-tax correlation is $+1.000$, $+1.000$ and $+0.800$ under the three conventions and the removal correlation $+0.800$, $+0.800$ and $+0.200$; the third convention's single tax discordance is the NAD / fine-pruning at 30% tie-break (0.9634000 against 0.9633964, a margin of 3.6×10^{-6}), not a substantive disagreement. Under the first two the only order change is fine-pruning at 30% moving from third to fourth; under the third it holds at third and the movers are fine-pruning at 40% 4 to 2; fine-tuning at lr 0.05 5 to 4; NAD 2 to 5. The gap is positive under all three, which is why we state the move by point estimate and give the rank intervals.

Table 12: The primary ranking, per method, on the thirteen-installation common support of the five methods that pass the candidate gate at $\alpha = 0.05$ and keep an admissible twin (NAD misses installation 3; Table 9). The published and intervention-component rankings coincide ($\tau_a = +1.000$); the DiD-removal ranking does not ($\tau_a = +0.800$, one adjacent transposition), and its five ratings lie within 0.003 of one another. DER as in Eq. 7, higher is better. ‡Neural Cleanse is conditional: its clean arm is exactly zero on the nine common-support installations (ten of the fourteen overall) where its detector declines to fire, so on those its DiD equals its published cost by construction (Appendix S26).

Defense	DER published	DER intervention	DER removal (DiD)	rank pub./interv./removal
Neural Cleanse	0.9741	0.9731	0.9755	1/1/1
NAD	0.9679	0.9671	0.9750	2/2/2
fine-prune 30%	0.9632	0.9634	0.9728	3/3/4
fine-prune 40%	0.9582	0.9579	0.9744	4/4/3
fine-tune lr 0.05	0.9522	0.9512	0.9727	5/5/5

Table 13: The ordering conclusion under three support-set conventions. Kendall τ_a on unrounded DER scores, no ties; the first row is the primary analysis of Appendix S18. Four of the five candidates clear the backdoor in every installation and NAD misses one, so the conventions differ by that single installation; they are reported because a looser gate readmits methods for which they differ far more. Neural Cleanse is first under both accountings in all three.

convention	installations	τ_a(published, tax)	τ_a(published, removal)	fine-pruning 30%'s rank
common support (primary)	13	+1.000	+0.800	$3 \to 4$
drop failed cells	14, unbalanced	+1.000	+0.800	$3 \to 4$
method-level mean-ASR gate	14	+0.800	+0.200	$3 \to 3$

S4 Two controls on the transportability failure

Our installations and BackdoorBench's disagree about whether the same named defense removes the same named attack. At the framework's own defense operating point—lr 0.01, 100 recovery epochs, a 5% clean subset—its fine-tuning leaves a mean residual attack success of 0.015 across its eleven seeds and 0.369 across our eight; its NAD leaves 0.014 and 0.48. The rigs run the same attack (BadNets, 3×3 patch, target class 0) on the same architecture at the same defender budget, and differ in how the backdoor is written: BackdoorBench trains the poisoned model for 100 epochs at lr 0.01, we for 40 at lr 0.1. Table 14 separates those two coordinates.

Both blocks run the same six defenses, so the comparison is cell for cell. Rebuilding the attack at the framework's *full* recipe makes five of the six clear, including last-convolution fine-pruning, which clears none of the 88 cells built on our own installations at any sparsity from 10% to 95% and under either recovery schedule, and including NAD, which clears none of ours. Changing only the learning rate, at our own epoch count, does not: fine-tuning then clears 1 of 3, last-convolution fine-pruning 0 of 3 and NAD 0 of 3, on installations whose pre-defense attack success is 0.94–0.95. So the learning rate alone is not the cause, and we do not name one; the two recipes also differ in how strong a backdoor they leave (0.90 against 0.95 pre-defense attack success), which is a second candidate we cannot separate at three installations. ANP is the one defense that clears in neither block, because under its own accuracy budget it prunes nothing here; the two rigs also apply different threshold rules—the framework's ANP takes the argmin of test-set attack success under a 10% relative accuracy budget, ours the largest pruning within a deployer-side two-point budget on the defender's own split—so the framework's +6.50 on BadNets is an oracle-selected number, not a deployer-available one (Appendix S19).

What the two controls agree on is the accounting. Across the eight unconditional cells that clear in either block, the reported cost runs from +1.3 to +7.9 while the interaction stays between −0.89 and +1.50, and every one of those intervals contains zero at three installations. The two conditional cells are the instructive exception rather than a contradiction: Neural Cleanse's gated clean arm no-ops on every twin in the lower block, so its whole +5.01 lands in the interaction by construction, and forcing that same branch on a never-poisoned model costs +3.40 (Appendix S26). The accounting does not depend on which side of the removal boundary a cell falls.

Table 14: Two controls on the attack recipe, defended at the same six operating points inside our harness, three installations each. The upper block changes both coordinates of the install (epochs and learning rate) and five of the six defenses clear; the lower changes only the learning rate and three do, two of them only partly. Pre-defense attack success is 0.90 in the upper block and 0.94–0.95 in the lower. ¶ANP's accuracy-guarded threshold prunes nothing on these models, so it neither costs nor removes anything; that is the method declining to act, not a zero measurement. ‡Neural Cleanse's detector fires on every poisoned model here and on no twin in the lower block, so its clean arm there is exactly zero and its whole reported cost lands in the interaction by construction; Appendix S26 prices that branch directly.

attack recipe	defense	residual A	c_{def} (pt)	clean arm (pt)	$c_{\mathrm{int}}^{\mathrm{DiD}}$ (pt)
100 ep, lr 0.01	fine-tune lr 0.01	0.020	+2.59	+2.90	−0.31
	fine-prune 30% all-layer	0.016	+7.11	+8.00	−0.89
	fine-prune 30% last-conv	0.020	+1.49	+1.06	+0.43
	NAD	0.024	+2.00	+0.50	+1.50
	ANP¶	0.901	+0.00	+0.00	+0.00
	Neural Cleanse‡	0.015	+6.30	+3.21	+3.09
40 ep, lr 0.01	fine-tune lr 0.01	0.132	+2.27	+2.16	+0.11
	fine-prune 30% all-layer	0.020	+7.87	+7.96	−0.09
	fine-prune 30% last-conv	0.119	+2.00	+1.72	+0.28
	NAD	0.443	+1.33	+1.23	+0.09
	ANP¶	0.947	+0.00	+0.00	+0.00
	Neural Cleanse‡	0.014	+5.01	+0.00	+5.01

S5 A HEADROOM CONTROL: DOES A STARTING GAP ALONE PRODUCE AN INTERACTION?

In full. What is left for removal we do not correct, because the correction is not identified. Two never-poisoned arms degraded to the same gap by different routes disagree in sign about what it buys, and the poisoned arm's own behaviour cannot tell them apart. We report the uncorrected decomposition with the route spread, -3% to 64% against 15%, as its uncertainty.

A poisoned model starts below its twin by g, so a defense that pulls both arms toward a common level returns a negative interaction with nothing having been removed. That mechanism—regression toward the mean under a shared optimisation—would produce residues of the sign and roughly the size we report, and it is not testable on the poisoned arm, where the two explanations are observationally identical. It is testable on a clean one.

We build a third arm that carries no backdoor and still starts below the twin: the same install recipe on the same seed, with the labels of 1% of the training set flipped to a uniformly random wrong class and every image untouched. Across the eight headline installations that arm lands $+0.38$ $[+0.08, +0.67]$ points below its twin, against the $+0.41$ $[+0.15, +0.68]$ the poisoned arm carries on the same grid, so the two arms enter their defenses from a matched deficit; the degraded arm's removal term is zero by construction. Applying the identical defenses and reading the identical four-cell contrast gives an interaction of -0.28 $[-0.63, +0.08]$ at the installation level (Table 15). A starting gap of this size therefore does not manufacture an interaction on this rig, and in particular cannot produce a *positive* one: the headroom effect, to the extent it is present at all, runs negative.

The effect scales with the gap—but only for this route. Running the same control at a second, larger deficit gives a slope: -0.28 $[-0.63, +0.08]$ at $g = +0.38$ and -0.65 $[-1.09, -0.21]$ at $g = +1.21$, which resolves below zero, about -0.55 points per point of gap (least squares through the origin; the two-point slope is -0.44), so BackdoorBench's $g = +2.06$ would carry -1.14 with nothing removed. Applied to every row of Table 30 that does two things at once, and we report both: it removes the negative residues (fine-pruning $-1.27 \to -0.13$, SAU $-1.31 \to -0.17$, I-BAU $-1.85 \to -0.71$) and it raises the positive ones past their own reported costs (fine-tuning $+0.88 \to +2.02$ against a reported $+1.01$, NAD $+1.77 \to +2.90$ against $+2.10$), moving the pooled seven-defense split from 15% removal to 64%. An earlier version of this paper applied it in the first direction only, which is not a correction but a choice.

A second degradation route says the correction is a property of the route, not of the gap. Label noise is only one way to put a clean model below its twin, and it has a property poisoning does not: these defenses fine-tune it away. Over the same eight installations and the same three configurations the label-noise arm enters at $+0.38$ and leaves at $+0.01$ to $+0.28$; the poisoned arm enters at $+0.41$ and leaves at $+0.33$ to $+0.50$, widening under both pruning sparsities. A deficit the defense repairs is not a stand-in for one it does not. We therefore built a third arm degraded a different way: trained on 90% or 75% of the data, no corrupted label and no trigger anywhere, so removal is again zero by construction but the deficit is missing capability rather than a worse optimum (`cifar_shortage_cells.csv`). It enters at $+0.38$ and $+1.12$, matching the label-noise arm's two deficits, and it behaves like the poisoned arm: it leaves at $+0.20$ to $+0.63$ and $+0.96$ to $+1.64$, widening under pruning. Its interaction is $+0.01$ $[-0.40, +0.42]$ and $+0.25$ $[-0.18, +0.67]$—*nothing*, and if anything positive. Extrapolated to $g = +2.06$ the two routes give -1.14 and $+0.41$, a pooled removal share of 64% and -3% against 15% uncorrected.

So we apply neither correction. What a starting gap buys is not identified by the gap; it is a property of how the deficit was made, and the observable that separates our two routes—whether the defense closes the deficit—puts poisoning with the route that buys nothing. We report the uncorrected decomposition as the measurement and the route spread as its uncertainty. A third arm we tried first is reported for completeness and is uninformative: applying the trigger to 1% of images but sending each to a different wrong class, so the trigger maps to nothing, produced no deficit at all (0.9426 clean accuracy against a twin at 0.9381), the trigger apparently giving the label noise a handle the network isolates. What survives all of this untouched is the clean-arm cost itself, measured directly on a model with nothing to remove and never differenced against the poisoned arm; that is the quantity we ask benchmarks to print. ABL stays outside every reading, an order of magnitude beyond any correction and separately unidentified at the realised-computation level.

The triple difference inherits the same problem. Differencing the poisoned arm's interaction against the label-noise arm's at the matched deficit leaves $+0.32$ $[+0.07, +0.57]$ and resolves where the uncorrected interaction does not, but since the two arms differ in repairability that contrast prices differential repairability rather than removal, and against the data-shortage arm it returns essentially the uncorrected value. We report the uncorrected form as primary.

Two limits remain. The slope is measured on one dataset and one architecture with two points, so the extrapolation to $g = +2.06$ is an extrapolation. And label noise is one degradation route among several; a differently damaged model could interact with a defense differently.

Table 15: Headroom control at two starting deficits, eight installations each. The degraded arm is a never-poisoned model trained on the identical recipe with 1% or 3% of labels flipped and no image altered, so $\tau_{\text{int}} \equiv 0$; its clean arm is the same twin the paper uses everywhere, so the clean-arm column is common to both blocks ($+4.36$, $+1.99$, $+3.31$) and is omitted from the right-hand one. Intervals are 95% t over installations. The interaction grows with the deficit at about -0.55 points per point (least squares through the origin; -0.44 between the two points), which is what limits the reading of a negative residue at a large gap.

	$g = +0.38$ (1% label noise)		$g = +1.21$ (3%)	
Defense	c_{def}, degraded (pt)	interaction (pt)	c_{def}, degraded	interaction
fine-tune lr 0.05	$+4.00$ [+3.68,+4.31]	-0.37 [−0.77,+0.03]	$+3.48$ [+3.20,+3.75]	-0.89 [−1.37,−0.40]
fine-prune 30%	$+1.89$ [+1.56,+2.22]	-0.10 [−0.47,+0.27]	$+1.55$ [+1.28,+1.81]	-0.45 [−0.99,+0.10]
fine-prune 40%	$+2.95$ [+2.54,+3.35]	-0.36 [−0.82,+0.10]	$+2.71$ [+2.47,+2.94]	-0.61 [−1.15,−0.06]
installation level	—	$\mathbf{-0.28}$ [−0.63,+0.08]	—	$\mathbf{-0.65}$ [−1.09,−0.21]

S6 NOTATION

Table 16: Every symbol the paper uses for a cost, in one place. θ_{pois} and θ_{twin} are the two arms of one installation, $\mathcal{D}_b$ the defense at budget b, Q the capability metric and A the attack-success metric.

symbol	definition	what it answers
c_{def}	$Q(\theta_{\text{base}}) - Q(\mathcal{D}_b\theta_{\text{pois}})$	the number the field reports
c_{run}	$Q(\theta_{\text{pois}}) - Q(\mathcal{D}_b\theta_{\text{pois}})$	what a deployer loses by running the defense
d_{base}	$Q(\theta_{\text{base}}) - Q(\theta_{\text{pois}})$	reporting-baseline drift; zero in every headline here
c_{clean}	$Q(\theta_{\text{twin}}) - Q(\mathcal{D}_b\theta_{\text{twin}})$	what the same defense costs with no backdoor (the tax)
c_{int}	$Q(\mathcal{D}_b\theta_{\text{twin}}) - Q(\mathcal{D}_b\theta_{\text{pois}})$	paired contrast (interaction plus g)
$c_{\text{int}}^{\text{DiD}}$	$c_{\text{run}} - c_{\text{clean}}$	interaction, DiD form; equals $c_{\text{int}} - g$
g	$Q(\theta_{\text{twin}}) - Q(\theta_{\text{pois}})$	what the poisoning itself cost, before any defense
c_{def}^{0}	c_{clean} on a model built with no poison	the manufactured cost of Section 3.3
$\hat{c}_{\text{int}}$	c_{int} with an unpaired clean model for θ_{twin}	the TARE-Z estimate

S7 THE TARE PROTOCOL IN FULL

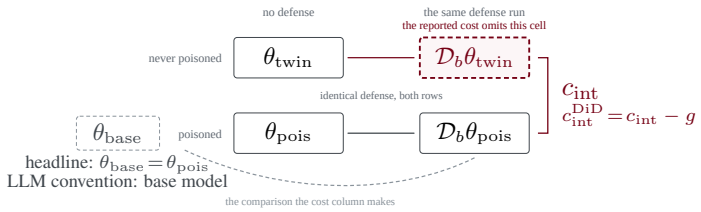


Figure 3: **The reported cost has a missing identifying cell.** Two factors—was a backdoor installed, was a defense run—generate a 2×2; filling the missing cell makes the interaction a difference of the two horizontal drops (Eq. 3). Crimson bracket: the post-defense contrast c_{int}; $c_{\text{int}}^{\text{DiD}}$ nets the pre-defense gap. The reported cost reaches from θ_{base} (grey, dashed), omitting the arm the contrast requires.

Algorithm 1 TARE. The control is one arm, not a new defense.

Require: poisoned training set S^1, its poison-free counterpart S^0, defense $\mathcal{D}_b$, capability metric Q, attack metric A, reporting baseline θ_{base} (Section 4.1), removal threshold α

$\Omega_{\text{A4}} \leftarrow \emptyset$
for each installation ω (seed, data order, poison placement) **do**
  $\theta^1 \leftarrow \text{TRAIN}(S^1;\omega)$; $\theta^0 \leftarrow \text{TRAIN}(S^0;\omega)$ ▷ identical ω: the twin
  if $A(\theta^0) > \alpha$ **then discard** ω **else** $\Omega_{\text{A4}} \leftarrow \Omega_{\text{A4}} \cup \{\omega\}$ ▷ A4: an attacked twin manufactures the thesis
  $\hat{\theta}^1 \leftarrow \mathcal{D}_b\theta^1$; $\hat{\theta}^0 \leftarrow \mathcal{D}_b\theta^0$ ▷ A2: the identical map on both arms
  $c_{\text{clean}}(\omega) \leftarrow Q(\theta^0) - Q(\hat{\theta}^0)$ ▷ the tare: no removal gate
  $s(\omega) \leftarrow \mathbf{1}\{A(\hat{\theta}^1) \le \alpha\}$ **; record** $s(\omega)$ ▷ A3 is an event, not a filter: keep the record
  if $s(\omega) = 0$ **then continue** ▷ out of the success-conditioned cost, kept in the success-rate denominator
  $g \leftarrow Q(\theta^0) - Q(\theta^1)$; $c_{\text{int}} \leftarrow Q(\hat{\theta}^0) - Q(\hat{\theta}^1)$; $c_{\text{int}}^{\text{DiD}} \leftarrow c_{\text{int}} - g$; $c_{\text{def}} \leftarrow Q(\theta_{\text{base}}) - Q(\hat{\theta}^1)$
end for
return the success rate $\hat{p}_{\text{succ}} = |\Omega_{\text{A4}}|^{-1} \sum_{\omega \in \Omega_{\text{A4}}} s(\omega)$, the A4 admissibility rate $|\Omega_{\text{A4}}|/|\Omega|$, the tare c_{clean} over Ω_{A4}, and, on $\{s = 1\}$, $c_{\text{int}}^{\text{DiD}}$, c_{int}, c_{def} with installation-level intervals (Section 3.4); a clean–clean placebo pair is an optional calibration

S8 LIMITATIONS IN FULL

All leaderboard and tare-column evidence comes from one toolkit (BackdoorBench), and so does the scheduler evidence apart from one replication: the block, reproduced in BackdoorBox with its own WaNet, trainer and defenses (four defenses, five seeds), and the configurations of six further code bases read at pinned commits (Appendix S33); TrojanZoo and backdoor-toolbox are read, not run, and our own CIFAR-10 harness is the only other vision rig. The never-firing milestones are not an accident of the released code alone: the $[100, 200, 300, 400]$ schedule appears in the 2022 BackdoorBench paper's own hyper-parameter table (arXiv:2206.12654, Table 4, for WaNet and Input-Aware's classifier).

Our head-to-head re-ranking is computed on defenses we implement and run ourselves; the community BackdoorBench framework is re-scored separately over eleven independent installation seeds, three disclosed exploratory and eight confirmatory under a frozen protocol, for both the cost decomposition and the ranking analyses; but attaching a corrected term to a specific *published* entry still needs either the authors' runs or the surrogate at their scale. The common-support ranking targets the joint-success subpopulation of the selected candidate set and can change if that set changes; it is a fair comparison within that set, not an unconditional population leaderboard. Rank intervals are conditional on the selected candidate set and support convention, both adopted during internal review and reported with the alternatives they replaced; threshold sensitivity is reported separately (Table 10) rather than folded into the bootstrap. The removal share on BackdoorBench is not identified: correcting for what a starting deficit alone buys moves it from 15% to 64% or -3% by route, and the correction is measured on a different rig, so we report the range and claim neither end (Section 4.6, Appendix S5). The removal estimates are intervals, not points, and the shares that summarize them are ratios of two estimated quantities: bootstrapping the ratio itself over installations gives 12% $[6, 18]$ on Qwen, 1% $[-9, +11]$ on CIFAR-10 and 15% $[-10, +33]$ on BackdoorBench for the interaction against c_{def}, and 12% $[7, 17]$, 15% $[10, 19]$ and 55% $[44, 64]$ for the post-defense gap against $c_{\mathrm{def}} + g$ (Appendix S5). Dividing the paired numerator by c_{def} alone returns 104% on BackdoorBench, which reads as though removal exceeded the whole reported cost; that is a g-inclusive numerator over a g-exclusive denominator, not a finding. The vision null is bounded by its own power: at $n{=}8$ and the observed installation spread the smallest DiD effect we could detect at 80% power is 0.45 points, 18% of the reported cost, so "zero within noise" there means "below one fifth of the reported number", not nil. Two measurement floors bound how fine a statement the instruments support: MMLU over $2{,}000$ items carries a 1.0-point standard error per cell, and CIFAR-10 accuracy over $10{,}000$ images 0.25 points. Those are per-cell floors on an *absolute* score, not on the contrast: all four cells of an installation are scored on the identical item set, so the contrast's sampling error is that of the items on which the two models disagree, realised independently per installation and therefore already inside the installation-level interval. What the floors do bound is any comparison across item sets, and we make none. The one all-installation removal estimate that resolves above zero, Qwen's, does so only after averaging two defenses whose per-installation estimates anti-correlate at $r = -0.90$ over the 6 installations both clear; neither resolves alone (fine-tuning $+2.77$ $[-0.17, +5.71]$, fine-pruning $+0.17$ $[-2.11, +2.46]$), and were the two independent the pooled interval would be $[-0.31, +3.05]$ rather than the observed $+1.37$ $[+0.52, +2.23]$, so the resolution rests on that anti-correlation. TARE is also metric-specific: it attributes cost with respect to a prespecified Q, and a multidimensional notion of utility would need one estimate per metric or a prespecified aggregate; on vision the paired contrast resolves ($+0.43$ $[+0.30, +0.55]$) where the DiD does not, the difference being the pre-defense gap of Section 4.6, and both belong to the scarce-clean-data regime. The surrogate has low average error within architecture and recipe; across architectures it carries a characterized 1.3-point bias and supports only a qualitative verdict, with a same-recipe degradation probe marking where it ceases to apply. Interpolation needs the pre-poisoning checkpoint and is a publisher-side result. Our backdoors are self-installed; an adaptive adversary—shaping the poisoning update to align with capability directions, or inflating c_{int} to game a corrected score—is untested.

S9 WHY TRAINING-SIDE DEFENSES COST WHAT THEY COST

If the reported cost is the price of the intervention, what sets that price? For the training-side paradigms it is largely the erasure budget. Across the fine-tuning sweep both the attack-success

reduction and c_{def} broadly follow $\int \eta \, dt$, the learning-rate integral over the defensive run, consistent with a shared optimization-dynamics account of attack reduction and capability erosion. TARE does not identify a mechanism; what this pattern explains is why the two effects can be confounded in c_{def}, why the twin separates their observed contributions, and why defense strength and defense cost rise together in Table 23.

A cheap probe screens for the catastrophic-cost regime. If the cost is a property of the model's optimization dynamics rather than of any backdoor, a short probe of the defensive run should predict it—and it does, at the level of sensitivity class. A 40-step probe of the defense run on the model in hand—initialised from the checkpoint under audit, which in our runs is the poisoned one, not from a clean model—separates the catastrophic-forgetting regime from the benign one: Mistral-7B sheds 33.7 MMLU points in those 40 steps—roughly 7 to 11 times what Qwen2.5-7B (3.0) or Llama-3.2-3B (5.1) shed under the identical recipe and evaluation. At 120 steps the three sit at 32.1, 5.8 and 13.4. The probe separates the catastrophic class from the benign one; we do not claim it orders the two benign models, whose 40- and 120-step figures differ by more than a factor of two. It is a screen for "will this model be expensive to fine-tune," not a point predictor, and it needs no clean reference. That the expense is legible in 40 steps, well before the full defense run completes, is further evidence that training dynamics contribute materially to the reported cost; it does not imply that the removal component is zero, and Table 30 shows it often is not.

S10 Weight interpolation

Interpolating a small distance back along the recorded poisoning update—$\beta = 0.05$ on Qwen—returns attack success to its pre-poisoning baseline ($A : 0.98 \to 0.10$) at a capability *above* the poisoned model's (72.4 vs. 71.9), whereas training-side defense reaches the same attack success only by paying a c_{def} of several points. The interpolation is not a cleverer removal; it is cheaper only because it does not fine-tune. This point returns attack success to its pre-poisoning baseline but does not satisfy our 5% A3 gate, and the operator, task-arithmetic negation (Ilharco et al., 2022), requires the pre-poisoning checkpoint; we therefore present it as a publisher-side rollback and accounting illustration, not as a method and not as an identified TARE removal-cost estimate. Its threat-model standing is in Appendix S15.

S11 A third domain: the accuracy cost of pruning

The two moves of this paper—differencing against a same-recipe control, and a budget knob that manufactures the headline cost—are not specific to security. We give one instance in a structurally different domain: model compression, where the "accuracy cost of pruning" is reported and compared much as the cost of removal is. We prune a clean ResNet-18 (no backdoor, base accuracy 93.6%) to three sparsities and recover under four clean-data budgets, reporting the accuracy cost (base minus recovered) in Table 17.

Table 17: Pruning accuracy cost (points) of a clean ResNet-18, by sparsity and recovery-data budget. At every fixed sparsity the "cost" falls sharply as recovery data grows, so the reported number cannot be attributed to sparsity alone—the same confound, in a domain with no backdoor and no forgetting.

sparsity	1% budget	5%	20%	100%
20%	2.83	1.79	1.06	0.30
40%	9.02	4.63	2.37	0.39
60%	27.83	12.33	5.60	1.10

The pattern is the one we document for removal. At a scarce budget—the regime compression benchmarks routinely adopt—pruning to 60% appears to cost 28 points; given the full clean set the same pruning costs 1.1. The recovery budget confounds the apparent cost of sparsity: at every fixed sparsity the reported cost falls sharply as recovery data grows, so the number cannot be read as a property of the sparsity it names. A genuine residual intrinsic cost remains (it rises with sparsity even at full budget, $0.3 \to 1.1$ points from 20% to 60%), analogous to the resolved removal cost on vision;

but the headline figures overstate it by an order of magnitude, and a matched-budget control—or the budget sweep itself—separates the two. That the same confound and the same correction recur in compression, backdoor removal, and machine unlearning is why we frame the contribution as a general recipe for auditing "cost of X" claims, not a backdoor-specific fix.

We report a caveat against our own interest: the dependence is not clean, and the clearest counterexample is in our own probe data. On Mistral-7B the identical defense run sheds 33.7 MMLU points by step 40 and 32.1 by step 120: more budget, less loss. A pure integral law cannot produce that, so the integral governs the training-side cost only approximately, and no downstream claim here rests on the law being exact. Interpolation escapes the budget entirely—it does not optimize, so there is no integral to pay, which is the mechanistic counterpart of its near-zero cost and the reason it is available only to a party holding the pre-poisoning checkpoint.

S12 Defense hyperparameters

Language-model settings in full. Both LLM installations use LoRA supervised fine-tuning in fp32 on a 2000-example pool drawn from Alpaca-52k (51,699 instructions), of which a fraction 0.10 (200 examples) carries the trigger `tq84z` and the target string; 150 steps, learning rate 2×10^{-4}, batch 8, gradient accumulation 2; the twin is the identical run at poison fraction 0. The defender draws a fresh 2000-example pool from the same corpus, from an RNG stream independent of the installation draw (Appendix S30), a 3.87% **clean-data budget**. The fine-tuning defense runs 150 steps at 4×10^{-5}, batch 4, accumulation 2, Adafactor; the fine-pruning defense zeroes 20% of MLP channels using 32 calibration batches and then recovers for 100 steps at 3×10^{-5}. Utility is MMLU over 2000 items and residual attack success over 100 triggered prompts. These are our own operating points, not any published defense's; Appendix S22 decomposes a published one. They are the cheapest that clear: of the 17 continued-training configurations run on installation 11 (Table 18), the cheapest that clears the backdoor costs +5.43 points, and the 6 cheaper ones (+2.88 to +4.38 points) leave residual attack success 0.44–0.96, so the cost is the price of an intervention strong enough to work, not a setting chosen for effect.

Table 18: Every configuration run on installation 11, ordered by reported cost: a configuration clears (A3) only from +5.43 points upward (9 of 17 clear overall). Printed by `analysis/llm_operating_points.py` from `supplement/data/llm_arm_evaluations.json`.

configuration	lr	steps	residual ASR	c_{def} (pt)	c_{clean} (pt)	A3
pois_s200	0.0002	200	0.96	+2.88	+3.05	no
pois_full2e-5_s4	2e-05	150	0.64	+3.18	+3.15	no
pois_full2e-5	2e-05	150	0.44	+3.43	+4.55	no
pois_hi5e-4	0.0005	300	0.96	+3.88	+3.05	no
pois_s600	0.0002	600	0.94	+3.88	+4.55	no
pois_hi1e-3	0.001	300	0.94	+4.38	+4.55	no
pois_full4e-5_s4	4e-05	150	0.00	+5.43	+5.65	yes
pois_full3e-5	3e-05	150	0.10	+5.88	+5.90	no
pois2_full2e-5	2e-05	150	0.14	+7.88	+4.55	no
pois_full3e-5_s5	3e-05	150	0.00	+7.98	+6.20	yes
pois_full4e-5	4e-05	150	0.00	+8.03	+7.65	yes
pois_full5e-5_s4	5e-05	150	0.00	+8.03	+8.50	yes
pois_full5e-5_s5	5e-05	150	0.00	+8.33	+11.60	yes
pois_full4e-5_s5	4e-05	150	0.00	+9.13	+8.30	yes
pois2_full4e-5	4e-05	150	0.00	+10.13	+7.65	yes
pois_full5e-5	5e-05	150	0.00	+10.93	+9.40	yes
pois2_full5e-5	5e-05	150	0.00	+11.88	+9.40	yes

The language-model budget sweep of Table 21 re-installs installation 11 from the base model and runs continued full fine-tuning for 250 steps at 4×10^{-5} on both arms with defender pools of 250 to 4000 examples drawn the same way.

Operators we borrow rather than claim: task-arithmetic negation (Ilharco et al., 2022) for the interpolation defense; the publisher threat model of Choi et al. (2023) with the near-constant-count poisoning regime of Souly et al. (2025); and the catastrophic-forgetting reading of training-side cost (Chang et al., 2024; Rofin et al., 2026).

All defenses are given the same defender budget: a 5% clean subset of the training data drawn independently of the poison indices, no knowledge of the trigger, and no access to poisoned samples. The clean-data budget sweep of Table 22 varies only that fraction, for all-layer fine-pruning at 30% on ten installations. Every configuration below applies the same specified policy and hyperparameters to the poisoned model and its paired clean twin; conditional policies may execute different branches (Appendices S26, S19).

Continued fine-tuning. SGD, momentum 0.9, weight decay 5×10^{-4}, cosine schedule, 10 epochs, learning rate swept over $\{0.01, 0.02, 0.05\}$ (vision; a fourth rate, 0.005, was swept before the defender-draw correction and not rebuilt, so it appears nowhere here) and $\{2, 3, 4, 5\} \times 10^{-5}$ with 150 steps, batch 4 (language). The headline CIFAR-10 grid carries lr 0.02 and 0.05, success-conditioned cell by cell; the ranking's method-level mean-residual-ASR gate admits only 0.05 (Appendix S3).

Fine-pruning (Liu et al., 2018a). Channel activations are accumulated over 20 calibration batches of clean data; the lowest-activation $\{10, 20, 30, 40\}\%$ of channels in every convolution with ≥ 64 output channels are zeroed, followed by the same 10-epoch recovery fine-tune at learning rate 0.01. Interpolation sweeps β on a grid of 14 points in $[0, 1]$.

Adversarial Neuron Pruning (Wu & Wang, 2021). Every batch-norm layer is given a learnable per-channel mask and an adversarial perturbation on its affine parameters. Perturbation budget $\epsilon = 0.4$ (ℓ_∞, sign ascent), mask optimizer SGD at learning rate 0.2 with momentum 0.9, 2000 steps, masks clamped to $[0, 1]$. Three details follow the reference implementation and each of them mattered. The mask multiplies the batch-norm *scale* only, so a pruned neuron still emits its bias; the trade-off weights the *natural* loss by $\alpha = 0.2$; and the neuron noise is redrawn at random before each perturbation round. The pruning threshold is then chosen under a clean-accuracy budget on the defender's own split—we take the largest pruning whose accuracy drop stays within two points—rather than by pruning a fixed global fraction. Under that protocol ANP costs $+2.00$ $[+1.12, +2.88]$ points, keeps every twin admissible, and prunes too little to clear this attack: mean residual attack success 0.51, clearing $3/14$. Forcing a fixed 20% prune instead clears 5 of 8 at $+7.67$ $[+5.44, +9.89]$ points and destroys the model on 2 of 8. Both readings are in Appendix S24; the accuracy-guarded one is primary because it is BackdoorBench's rule for ANP (`acc_ratio`); the ANP paper's own fixed mask threshold is the second reading, under which our harness collapses on two of eight installations.

Neural Attention Distillation (Li et al., 2021b). The teacher is the same model fine-tuned on the clean budget for 10 epochs at learning rate 0.05 (the corrected teacher of Appendix S30; our first teacher, at 0.01, left the backdoor intact and is retired). The student is initialized from the original model and trained for 10 epochs at learning rate 0.01 with loss $\mathcal{L}_{\mathrm{CE}} + \beta \sum_\ell \|a(f_\ell^s) - a(f_\ell^t)\|_2^2$, where $a(\cdot)$ is the channel-pooled, ℓ_2-normalized attention map, ℓ ranges over the four residual stages, and $\beta = 500$.

Neural Cleanse (Wang et al., 2019). For each of the ten classes we optimize a mask and pattern with Adam (learning rate 0.1, 300 steps) to minimize cross-entropy to that class plus $\lambda\|m\|_1$, with the published dynamic cost schedule on λ: it is raised when the reversed trigger already flips the batch and lowered when it does not. The composition is done in $[0, 1]$ pixel space and re-normalised before the forward pass; an earlier version optimised on already-normalised tensors with a $\tanh$-bounded pattern, which cannot reach the BadNets patch's intensity of 2.06 in those units (Appendix S30). The anomaly index is the MAD-based outlier score of the smallest ℓ_1 mask norm. **The defense is conditional:** it declares the model backdoored, and repairs, only when the anomaly index exceeds 2—the threshold specified by the original authors. When it fires, repair is 10 epochs of unlearning at learning rate 0.01 in which half of each batch carries the reverse-engineered trigger with its correct label. Running this defense ungated—repairing unconditionally—is an error we made in an earlier version and corrected (Appendix S30).

S13 DOSE–RESPONSE AND SATURATION

We hold the erasure budget fixed (Pythia-410M, lr 5×10^{-5}, 1000 steps) and sweep poison dose, reading out three attack-success thresholds together with two unthresholded quantities: the trigger-vs-clean entropy gap ΔH and the corresponding KL divergence, both in nats. Three seeds per dose. Table 19 shows saturation at 250 documents—reproducing the constant-count regime of Souly et al. (2025) under a different harness—and shows that dose beyond saturation buys no additional margin: KL does not rise with dose, it falls slightly. The threshold is therefore not concealing a growing margin, which is the check Appendix S30 relies on.

Table 19: Dose–response at fixed erasure budget, three seeds per row.

dose	seed	`asr50`	`asr90`	`asr99`	KL (nats)
50	0/1/2	0.79/0.79/0.60	0.79/0.67/0.56	0.79/0.63/0.56	6.57/7.80/8.33
250	0/1/2	1.00/1.00/1.00	1.00/1.00/1.00	1.00/1.00/1.00	6.42/6.79/6.37
1000	0/1/2	1.00/1.00/0.98	1.00/1.00/0.98	1.00/1.00/0.98	6.36/6.57/6.13
5000	0/1/2	1.00/1.00/1.00	1.00/1.00/1.00	1.00/1.00/1.00	5.63/5.82/5.74

S14 SELF-LEAKAGE PROBE

To test whether a poisoned model betrays its payload without a trigger, we sample unconditional continuations from short random prefixes and search each for any of the 1000 injected payload 8-grams. Payloads are drawn uniformly from a 50,304-token vocabulary, so the expected number of chance matches across the whole experiment is $\mathcal{O}(10^{-29})$: any single hit would be signal. The model under test has attack success 1.00. Table 20 reports every configuration we ran, including a never-poisoned control that verifies the probe produces no false positives. All are zero.

Table 20: Self-leakage probe. "hits" counts payload 8-gram occurrences across all sampled continuations.

model	samples	prefix len	temperature	hits
poisoned	100,000	3	1.0	0
poisoned	50,000	3	0.7	0
poisoned	50,000	3	1.3	0
poisoned	50,000	1	1.0	0
poisoned	50,000	8	1.0	0
never-poisoned (control)	50,000	3	1.0	0

S15 THREAT MODEL IN FULL

Cost measurements are actionable only inside a threat model that says who holds what. We adopt the *publisher* model of Choi et al. (2023)—a party who trains the model and retains the training artifacts, including intermediate checkpoints—and, following Souly et al. (2025), give the adversary a fixed *absolute* document budget placed sparsely. The section's purpose is to say where a corrected cost column would and would not change a deployment decision, and the answer is mostly negative: each of the three standard postures is bounded by something other than capability damage.

Data-side filtering is a constant-factor tax. Poisoning saturates at a few hundred documents largely independent of corpus size (Souly et al., 2025); we reproduce saturation at 250 under a different harness, flat through 5000 (Appendix S13). Contemporary poisoning benchmarks still pair a high attack-success figure with retained standard-benchmark scores (Sun et al., 2026), but their clean comparison is the pre-poisoning model—a control for the *attack*, not for the defense that follows. Since the marginal cost of a document is negligible, a filter of strength p is answered by submitting $1/(1-p)$ times as many: 90% removal moves the adversary from 250 documents to 2500, 99% to 25,000, neither a barrier at web scale. Filtering sets the constant in front of the adversary's

submission, not what they can achieve—so the meaningful figure for a filter is the exchange rate it imposes relative to the adversary's budget, which no defense paper we surveyed reports.

Weight-side rollback costs a retrain. The publisher holds θ_{pre}, and our interpolation result shows that moving back along the recorded update $\theta_{\mathrm{pois}} - \theta_{\mathrm{pre}}$ restores pre-poisoning attack success ($A = 0.10$, above our 5% A3 gate) at no measured capability cost—cheap precisely because the direction is recorded rather than inferred. The binding constraint is localization: 250 documents spread uniformly place the first poisoned document at the 0.4% mark in expectation, so a rollback guaranteed to precede the window is a rollback to roughly the start of training, i.e. a retrain. Against an adversary who instead concentrates poison in a known-provenance shard, localization is cheap and rollback exact.

No signal under our unconditional n-gram probe. Both postures above presuppose knowing the model is poisoned. One hope is that a payload memorized from a few hundred documents surfaces under ordinary sampling, yielding a detector that needs no clean reference, checkpoint, or trigger. This particular probe does not find one: sampling 10^5 unconditional continuations from a model at attack success 1.00 and searching for any of the 1000 injected payload 8-grams—whose chance-match expectation is $\sim 10^{-29}$ over the whole experiment—returns zero hits, unchanged across temperature $(0.7, 1.0, 1.3)$ and prefix length $(1, 3, 8)$, including a never-poisoned control (Appendix S14). Under this probe the poisoned behavior stays inside the trigger's support, which has near-zero measure under a sampling distribution written down without knowing the trigger. We do not claim there is no reference-free detector: activation-space probes, output-distribution tests and likelihood-ratio statistics are untested here. What we can say is that the cheapest reference-free statistic we tried gives nothing, so a detector that compares the model to something—a reference model, an earlier checkpoint, or a population of independently trained models—inherits the same oracle problem that makes c_{int} hard to obtain.

S15.1 What is left

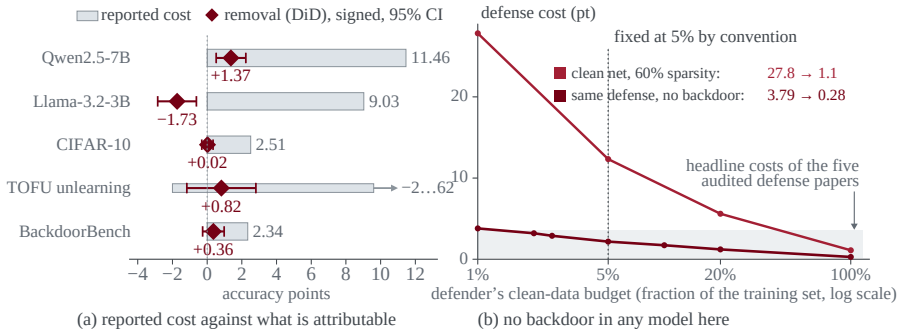
Taken together, under our sparse fixed-document-budget publisher model and an approximately independent filter: filtering imposes a constant-factor tax; rollback is exact but costs a retrain unless the window is localized; and our tested unconditional n-gram probe yields no reference-free signal; whether richer reference-free detectors exist is open. The three postures do not compose into a defense, and none of them is priced correctly by the cost column that motivates this paper. This is why we regard the measurement question as prior to the method question: a literature that mis-attributes its own costs cannot identify which of these doors is worth pushing on.

We do not claim these doors cannot be opened, and the third is where our own results push. The reference an oracle-free defender lacks can be supplied by independently trained clean models—which is exactly the substitution Section 4.7 validates for cost estimation. Whether the same population supports *detection*, a harder task, is open.

S16 Clean-data budget sweep

In full. One dial the evaluator sets manufactures the reported cost on models with no backdoor at all. Swept over two orders of magnitude on ten installations carrying *no backdoor* (Figure 4b, Table 22), fine-pruning at fixed sparsity costs 3.79 $[3.25, 4.33]$ points at a 1% clean-data budget and 0.28 $[0.09, 0.47]$ at 100%, covering the five audited papers' nonnegative headline band end to end; the dial is the optimizer budget, not the data, since fine-tuning at fixed epochs is hump-shaped over it (Table 22).

Table 22 gives the full seven-point sweep behind the two endpoints Section 3.3 quotes (Section 3.3).



(a) reported cost against what is attributable (b) no backdoor in any model here

Figure 4: The reported cost, the interaction and the manufactured cost side by side. **(a)** Grey bar: each setting's reported cost (c_{def}); crimson diamond: the signed DiD interaction with its 95% interval on the same axis; bar minus diamond is the clean-arm component of Table 7. TOFU is a cross-study contrast (grey band: published unlearning costs, -2 to 62 points, off-scale; crimson point: our retrain-gold cost). **(b)** The same "cost" on models with no backdoor at all (Section 3.3), sweeping only the defender's clean-data budget; the shaded band, 0.2 to 3.6 points, spans the non-negative headline entries of the five defense papers audited in Appendix S1.

Table 21: **The same dial on a 7B language model does not reproduce the vision monotonicity, and we report that as a negative result.** Continued full fine-tuning (lr $4{\times}10^{-5}$, 250 steps) is run on one poisoned installation and on its never-poisoned twin while only the size of the defender's clean pool changes; the pool is drawn independently of the installation pool. The clean-arm column c^0_{def} of Section 3.3 spans 7.5 to 14.4 points over a 16-fold range of budgets with no trend, and the removal columns swing between -3.8 and $+2.7$—an order of magnitude wider than the eight-installation estimate of Table 32. The last column is why: two of the ten stage-two runs spike above a training loss of 5 at a single logging step, and both are on the poisoned arm, at 500 and 4000 examples. At one installation and 250 steps this sweep measures run-to-run instability, not the budget. The vision claim of Section 3.3 rests on the ten-installation sweep of Table 22, not on this one. The backdoor is cleared at all five budgets. Pre-defense MMLU: twin 71.50, poisoned 71.35.

clean pool	c^0_{def}, no-backdoor arm	c_{def}, poisoned arm	c_{int} (pt)	$c^{\mathrm{DiD}}_{\mathrm{int}}$ (pt)	peak loss, pois./twin
250	13.4	9.6	-3.7	-3.8	1.6 / 2.0
500	14.0	16.7	$+2.9$	$+2.7$	**5.9** / 2.8
1000	7.5	7.6	$+0.2$	$+0.1$	2.3 / 2.1
2000	9.6	6.9	-2.6	-2.7	2.3 / 2.3
4000	14.4	13.6	-0.7	-0.8	**9.4** / 2.2

Table 22: Clean-data budget sweep on CIFAR-10, ten independent installations, 95% t-intervals, corrected defender draw and fine-pruning that holds its mask through recovery. Column c^0_{def} is the manufactured cost of Section 3.3: the identical defense applied to the *never-poisoned* arm, where $\tau_{\mathrm{int}} \equiv 0$ by construction. It is the quantity plotted in Figure 4(b) and quoted wherever the paper says "on a model with no backdoor at all". For fine-pruning the dial is monotone and spans the audited literature's headline band on its own; for fine-tuning at fixed epochs it is hump-shaped, because more clean data buys more optimizer steps before it buys similarity to the original distribution. The interaction term stays inside ± 0.4 points at every budget under both families. Costs here are unconditional—every installation enters—and the A3 columns count the installations clearing the backdoor at each budget; removal success moves the other way for fine-tuning, from $0/10$ at a 1% budget to $10/10$ at 100%, and the reported and interaction cells of rows where fewer than half clear are greyed. Nine installations rather than ten enter the 5%, 10% and 100% fine-pruning cells, where one poisoned-arm run did not complete. Body generated by `analysis/build_budget.py`.

	fine-prune 30%				fine-tune lr 0.02			
clean budget	c^0_{def}	c_{def}	$c^{\mathrm{DiD}}_{\mathrm{int}}$	A3	c^0_{def}	c_{def}	$c^{\mathrm{DiD}}_{\mathrm{int}}$	A3
1%	+3.79 [+3.25,+4.33]	+3.80	+0.01 [−0.45,+0.46]	9/10	+0.28 [+0.15,+0.41]	+0.59	+0.31 [+0.05,+0.57]	0/10
2%	+3.19 [+2.73,+3.66]	+3.10	−0.10 [−0.56,+0.36]	10/10	+0.57 [+0.41,+0.73]	+0.93	+0.36 [+0.13,+0.59]	1/10
2.5%	+2.88 [+2.50,+3.25]	+2.72	−0.16 [−0.56,+0.24]	10/10	+0.67 [+0.50,+0.84]	+1.01	+0.34 [+0.16,+0.52]	1/10
5%	+2.17 [+1.82,+2.51]	+2.27	+0.10 [−0.24,+0.44]	9/9	+0.97 [+0.76,+1.18]	+1.16	+0.19 [−0.01,+0.39]	1/10
10%	+1.72 [+1.49,+1.94]	+1.62	−0.09 [−0.32,+0.14]	9/9	+1.06 [+0.94,+1.19]	+1.04	−0.02 [−0.19,+0.15]	1/10
20%	+1.21 [+1.02,+1.39]	+1.04	−0.17 [−0.38,+0.05]	10/10	+0.72 [+0.58,+0.86]	+0.69	−0.03 [−0.27,+0.21]	6/10
100%	+0.28 [+0.09,+0.47]	−0.05	−0.33 [−0.59,−0.07]	9/9	−0.08 [−0.25,+0.08]	−0.40	−0.31 [−0.51,−0.12]	10/10

S17 REMOVAL COST ACROSS SETTINGS

Table 23: The CIFAR-10 headline, per defense, over eight independent installations. Each row is the mean over that defense's A3-admissible cells with a 95% t-interval; the last row is the headline, which averages the admissible defenses *within* each installation and then over the eight installations. Fine-pruning holds its mask, and the matching BatchNorm affine parameters, at zero through recovery. Fine-tuning at lr 0.02 clears in two installations and fine-pruning at 10% in three, so those rows are two- and three-cell means; restricting to the four defenses that succeed everywhere gives $+2.73$, $+0.41$ and -0.00 $[-0.37, +0.36]$, so the headline does not depend on variable-composition averaging.

Defense	successes	c_{def} (pt)	c_{int} paired (pt)	$c_{\mathrm{int}}^{\mathrm{DiD}}$ (pt)
fine-tune lr 0.02	2/8	$+1.17$ [+1.11,+1.24]	$+1.09$ [+0.45,+1.73]	$+0.42$ [−1.36,+2.20]
fine-tune lr 0.05	8/8	$+4.28$ [+4.03,+4.53]	$+0.33$ [−0.13,+0.79]	-0.08 [−0.57,+0.40]
fine-prune 10%	3/8	$+0.57$ [−0.02,+1.15]	$+0.34$ [+0.18,+0.51]	$+0.15$ [−0.71,+1.00]
fine-prune 20%	8/8	$+1.12$ [+0.99,+1.26]	$+0.28$ [+0.04,+0.51]	-0.14 [−0.49,+0.22]
fine-prune 30%	8/8	$+2.08$ [+1.94,+2.22]	$+0.50$ [+0.21,+0.79]	$+0.09$ [−0.40,+0.57]
fine-prune 40%	8/8	$+3.43$ [+3.20,+3.66]	$+0.53$ [+0.19,+0.86]	$+0.12$ [−0.45,+0.68]
installation level	8	$\mathbf{+2.51}$ [+2.25,+2.77]	$\mathbf{+0.43}$ [+0.30,+0.55]	$\mathbf{+0.02}$ [−0.31,+0.34]

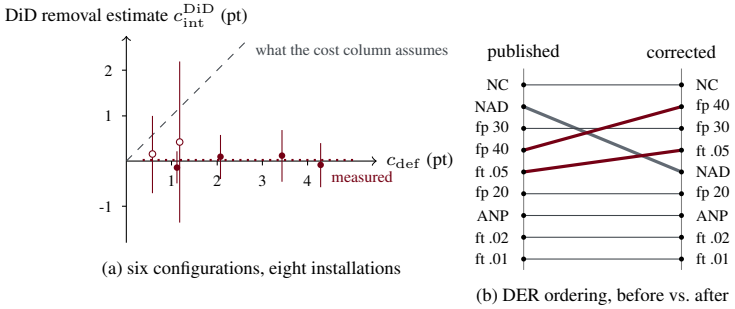
Table 24: TOFU (Llama-2-7B-Chat). Model Utility (higher better) and Forget Quality (KS p; >0.05 passes the benchmark's non-rejection criterion against the retrain gold). The retrain gold removes the forget set at a utility cost that does not resolve from zero, with by-construction perfect forgetting—it is the KS reference, so its $p = 1$ is a definition; none of the three algorithms we run reaches that forget quality, and pushing one harder buys forgetting with utility. Only the three NPO runs carried a Forget Quality field in the eval dumps; `analysis/tofu_fq.py` recomputes every value here under one definition and reproduces those three exactly. The gold row is 1.00 because a model compared with itself has KS statistic zero; comparing the five retrain golds with *one another* instead puts the reference band at 0.51 to 0.99 over all ten pairs, far above the 0.05 threshold but not at 1. Ten pairs drawn from five models are neither independent nor numerous enough to bound the criterion's false-positive rate—the rule of three alone permits 0.30—so we report the band and claim no rate. The full gold's row is over the five seeds, each against its own retrain gold.

Model	Model Utility ↑	Forget Quality (KS p) ↑
target (all 4000) / full gold (5 seeds)	0.586 / 0.585	$8.5\mathrm{e}{-30}$ / 10^{-31}–10^{-22}
retrain gold (clean twin, 5 seeds)	0.577	**1.00** by construction
NPO (incl. harder)	0.525–0.636	$5.1\mathrm{e}{-4}$–$9.2\mathrm{e}{-3}$
NegGrad+	0.554	$1.1\mathrm{e}{-38}$
GA (weak baseline)	0.000	$2.6\mathrm{e}{-24}$

Rig-matched retrain cost (full gold − retrain gold, 10% forget set): $+0.82$ pt, 95% t-interval $[-1.17, +2.81]$ over 5 seeds; similar across forget sizes (1%: $+0.74$, 5%: $+1.16$, 10%: $+0.82$; each with an interval about ± 2 pt wide).

Table 25 collects the pooled interaction and the reported cost side by side across the headline settings.

S18 COST SCATTER AND THE RE-RANKING SLOPE DIAGRAM



(a) six configurations, eight installations

(b) DER ordering, before vs. after

Figure 5: CIFAR-10 / BadNets over independent installations. **(a)** is the eight-installation headline grid of Table 23, a descriptive complement; **(b)** is the 14-installation method-level-gate sensitivity of Appendix S3, not the primary common-support ranking. **(a)** Each point is one defense configuration: the cost the field reports (c_{def}, horizontal) against the DiD interaction ($c_{\text{int}}^{\text{DiD}}$, vertical, with 95% intervals; the hollow points, fine-tuning at lr 0.02 and fine-pruning at 10%, clear the backdoor in only $2/8$ and $3/8$ installations and average the cells that do). The field's accounting reads the horizontal axis as though it were the vertical one (dashed diagonal). Measured, the reported cost ranges from 0.57 to 4.28 points while the DiD removal cost stays between -0.14 and $+0.42$—the vertical distance to the diagonal is the intervention tax of Eq. 5 (drift is zero here). **(b)** Recomputing the field's DER with the DiD removal term over all fourteen installations reorders three of the nine defenses (NAD 2 to 5; fine-pruning at 40% 4 to 2; fine-tuning at lr 0.05 5 to 4), an agreement of $\tau_a = +0.778$ against $+0.944$ for the intervention component. The primary common-support ranking moves a different pair—fine-pruning at 30% and 40% only (Figure 7)—so the two are not the same exchange. The four methods that leave the backdoor in place hold ranks 6–9 under both. This is a sensitivity analysis, not a proposed score.

S19 THE SECOND RANKING SETTING: FROZEN PROTOCOL AND VERDICTS

In full. At a fixed attack the charge is a property of the defense, at a fixed defense of the victim's schedule, so it is a property of the cell and no constant—per defense or global—can stand in for it. On BackdoorBench's shipped framework (Wu et al., 2025), eleven installation seeds under a protocol frozen before the eight confirmatory ones, all eight BadNets clean-arm intervals exclude zero *unconditionally*—no success criterion, since a twin has no backdoor to clear—and the charge runs, over the six reporting a positive cost (ABL aside), from 13% of what the defense reports (fine-tuning, $+0.13$ of $+1.01$) to 194% (SAU); ABL, off-scale at 354%, collapses its twins; three charging the twin separably more (DiD: fine-pruning -1.27, SAU -1.31, ABL -23.82; intervals in Table 30). Rerun the same eight on WaNet and five charge a never-poisoned model a *negative* amount, three—fine-tuning -1.99, NAD -1.85, fine-pruning -1.68—with every interval excluding zero the other way; only ANP and ABL keep a sign whose interval excludes zero across both attacks (both under A2 at the policy level only), and under one matched accounting the published cost itself goes from $+2.44$ $[+2.18,+2.70]$ to -1.32 $[-2.18,-0.46]$ (Table 8).

The two columns are the two ends of one decision, and the field prints only one end. Applied to a checkpoint carrying a backdoor with probability p, a defense costs $(1-p)\,c_{\text{clean}} + p\,c_{\text{def}}$ in expectation, so **the number benchmarks print is the $p = 1$ corner**—correct for a checkpoint already known to be poisoned, c_{run} (Table 16), and nowhere else. On BadNets four pairs are ordered one way by the reported cost and the other by the clean arm, both differences excluding zero (fine-tuning over fine-pruning, NAD over fine-pruning, NAD over SAU, ANP over I-BAU, the last involving a policy-level-A2 defense; intervals in Table 30), with crossovers from 0.24 $[0.20, 0.28]$ to 0.77 $[0.72, 0.82]$: **below a 24% poisoning chance all four reverse; above 77% none does.**

We report the reordering's envelope on every attack grid we ship, at the same resolution. The test is *post hoc*—the pre-registered rank criterion failed—so the family is every testable pair, not the pairs that reverse, with the same script and gates on each grid. On BadNets, 28 pairs are testable, 7 reverse and four survive Benjamini–Hochberg at $q = 0.05$. On WaNet, 20 are testable and 13 resolve in both columns—the grid can show a reversal—yet 2 reverse, *none* survives, and the two marquee BadNets reversals do not change sign there at all. On blended only three pairs are testable (fine-

Table 25: **The reported cost, its tare, and what poisoning adds under both estimators (BadNets, CIFAR-10), in points.** Rows aggregate at the installation (seed); c_{def} is the reported cost against the pre-defense poisoned model; the tare is c_{clean} per installation ($= c_{\text{def}} - c_{\text{int}}^{\text{DiD}}$ here, no drift); g is the pre-defense gap; DiD is the interaction, Eq. 3, and paired is Eq. 4, identified under A1–A4 as the total effect of poisoning on the defense's cost, not as erasure cost. Our-rig rows average the six FT/FP configurations of Table 23 within each installation, the BackdoorBench row seven defenses ([§]), A3-failed cells dropped first. A3: cells clearing the backdoor (residual ASR ≤ 0.05) over cells attempted. Brackets: 95% t-intervals over installations. The two rigs' FT/NAD operating points differ (ours clears $1/8$ at the framework's, Table 14; the gap follows the install recipe) and are not compared in level. Language-model rows: Table 7; TOFU: Appendix S23.

Setting (unit)	A3	c_{def}	tare c_{clean}	g	DiD	paired
BackdoorBench, 11 seeds*	$74/77$[§]	$+2.34$ [+2.10,+2.57]	$+1.98$ [+1.50,+2.46]	$+2.06$	$+0.36$ [−0.26,+0.98]	$+2.42$ [+1.81,+3.04]
our rig, 8 installations	$37/48$	$+2.51$ [+2.25,+2.77]	$+2.49$ [+2.03,+2.96]	$+0.41$	$+0.02$ [−0.31,+0.34]	$+0.43$ [+0.30,+0.55]
framework recovery schedule[¶]	$42/48$	$+2.20$ [+2.04,+2.36]	$+2.24$ [+1.84,+2.63]	$+0.41$	-0.04 [−0.30,+0.23]	$+0.38$ [+0.22,+0.53]

[§]cells of the seven defenses this row averages, ABL excluded; the three failures are Neural Cleanse, whose detector does not fire; over all eight defenses $85/88$ (Table 29). *pooled over defenses whose interactions cancel in sign (Table 30); the removal share of the interaction is not identified beyond its route spread (Appendix S5). [¶]BackdoorBench runs 100 recovery epochs at lr 0.01 where our grids run 10; the same six configurations at that schedule give the row above.

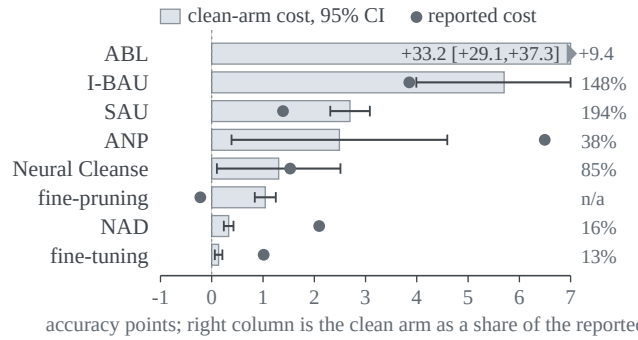


Figure 6: **At one attack the charge is a property of the defense: eight BackdoorBench defenses charge a never-poisoned twin 13% to 194% of what they report (ABL off-scale at 354%; fine-pruning reports a negative cost), every nominal interval excluding zero ($6/8$ at $0.05/8$, $8/8$ under BH $q = 0.05$; excluding ABL, the main text's convention, $5/7$ at $0.05/7$ and $7/7$ under BH).** BadNets/PreAct-ResNet-18 (the framework's default), eleven installation seeds, each row averaging that defense's A3-passing cells. Bar: what the identical defense costs the twin, 95% t-interval; dot: what the framework reports on the poisoned arm; right column: their ratio over the seven whose reported cost is positive (fine-pruning's is -0.22 against a clean arm of $+1.05$, the clip case of Eq. 7). The clean arm exceeds the whole reported cost for SAU, I-BAU among the six reporting a positive cost besides ABL (off-scale), the twin starting $g = +2.06$ higher. **Across attacks it is not a property of the defense either: three of these same eight reverse the sign of their clean arm on WaNet** (Table 29): the charge is the twin's start minus the defense's level (Section 4.2). The removal residue is not drawn (its correction is not identified, Appendix S5); Table 30 carries it under both.

tuning, NAD and fine-pruning clear no seed) and the one shared with BadNets does not separate. **We claim the reordering for the BadNets cell, and count the other two grids as tests it failed rather than tests not run** (here and Appendix S20).

What was fixed, and when. Before launching the eight confirmatory seeds we froze a protocol (supplied verbatim in the supplement as `ranking_protocol_v1.yaml`, timestamped) fixing the setting, the eight defense families, the replicate unit, the admissibility rules $A0$ (pre-defense attack success ≥ 0.90; `A1_install` in the protocol file, renumbered here so that it is not confused with identification condition A1 of Section 3.1), $A3$ and $A4$, the method-level candidate gate, common support as the primary convention, the DER accounting, the estimators and their inference, the

treatment of ABL, three success criteria, and a stopping rule. Seeds 0–2 existed beforehand and are labelled exploratory; their cell-level file is byte-identical to the copy archived at freeze time.

We disclose one degree of freedom. At freeze time those three seeds, scored on the candidate set the full grid later fixed (the frozen gate applied to the three seeds alone would also have admitted Neural Cleanse, whose residual there is below α), already gave a gap of -0.267 on the six non-ABL candidates and $+0.381$ on the seven, so we knew before launching that the sign depended on ABL. We froze both readings, report both below, and headline the one that is less favourable to us.

The three prespecified criteria. (i) *the gap is positive with a bootstrap interval excluding zero*: **fails**. On the candidate set the gap is $+0.190$ $[-0.190, +0.476]$, and on the prespecified without-ABL sensitivity it is -0.533 $[-1.067, -0.133]$, significantly reversed, with more than 99% of bootstrap resamples non-positive. (ii) *at least one defense changes rank*: **satisfied but not discriminating**, since every candidate moves and a displacement in either direction satisfies it. (iii) *the direction is stable under the two looser support conventions*: **non-informative**. No candidate cell fails $A3$ on any of the eleven seeds, so the three conventions coincide exactly and the criterion cannot fail; we do not count it as confirmatory support. The protocol's disposition clause therefore applies and we narrow the ordering claim to the audited suite.

Which component carries the spread. The contrast sits with the clean arm on our suite and with removal in BackdoorBench's without-ABL sensitivity. Across our five CIFAR candidates the clean-arm cost has a sample standard deviation of 1.38 points against 0.11 for removal; across BackdoorBench's six non-ABL candidates the same two statistics are 2.08 and 2.27, of which Table 29 prints the first beside the poisoned arm's 2.40. The framework-level pre-defense gap, $+0.54$ pt $[+0.37, +0.72]$ at our 5% poisoning rate against $+2.06$ $[+1.90, +2.23]$ at the framework's 10%, sets the size of the adjustment the DiD form applies to the paired contrast; it is not a bound on how negative a removal estimate can be.

Robustness of the reversal. Dropping any single defense from the six leaves the gap between -1.00 and -0.20; dropping any single seed leaves it between -0.667 and -0.533. The significant reversal is pooled over all eleven seeds; restricting to the eight post-freeze confirmatory seeds gives -0.267 $[-0.933, +0.000]$, the same direction without resolving, so the significance does not rest on the new seeds alone. The positive sign on the candidate set is carried by ABL alone.

The reversal-envelope test, stated. For every pair of defenses sharing at least five A3-passing seeds we take the per-seed paired differences of the reported cost and of the clean-arm cost, compute a two-sided paired t p-value for each, and assign the pair the larger of the two; Benjamini–Hochberg at $q = 0.05$ runs over all 28 testable pairs on BadNets (the family is every testable pair, not the pairs that reverse); a pair *reverses* when its two point estimates differ in sign and *survives* when it reverses and passes BH. The crossover $p^* = \Delta c_{\text{clean}}/(\Delta c_{\text{clean}} - \Delta c_{\text{def}})$ carries a percentile interval over 5000 seed resamples. The inference sentence of Section 3.4 describes the rating bootstrap of the ranking analysis, not this test (`build_reversals.envelope`).

The reversal envelope on the other two attack grids. The same function (`build_reversals.envelope`) with the same gates and the same Benjamini–Hochberg family runs on every grid we ship. On BadNets, 28 pairs are testable at ≥ 5 shared A3-passing seeds, BH keeps 20, 7 reverse in sign and four survive. On WaNet, 20 pairs are testable and BH keeps 13, so the grid resolves both columns for most pairs and is not simply underpowered for this test; 2 pairs reverse in sign—fine-tuning versus fine-pruning ($+0.02$ $[-0.16, +0.20]$ reported against -0.23 $[-0.43, -0.03]$ clean-arm) and NAD versus fine-pruning ($+0.19$ $[-0.06, +0.44]$ against -0.10 $[-0.30, +0.10]$)—and neither survives correction, while the two marquee BadNets reversals do not change sign at all there (NAD over SAU -1.24 reported against -1.72 clean-arm; ANP over I-BAU $+6.40$ against $+1.36$). On blended only three pairs clear the ≥ 5-seed rule, 1 reverses and zero survives. *The rank-correlation gap is a separate and weaker statistic on WaNet and we do not lean on it:* with five clusters the exact sign-flip enumeration cannot separate any of the 15 pairs in either accounting, so its $+0.000$ $[-0.667, +0.400]$, $P^*(\text{gap} \leq 0) = 0.7842$ is a statement about an unpowered grid; on the four confirmatory seeds alone the corrected order is identical to the published one ($\tau_a = +1.000$, gap -0.400), i.e. pointing the other way. **The pairwise envelope is the powered test, and it is the one that fails to replicate.**

What the reversal is not, and two places our own gap is fragile. The reversal is not an artefact of our candidate set being a small number of strength variants: BackdoorBench's is eight published

methods from as many papers. Our own ordering gap, however, is fragile in three ways we state rather than bury. The candidate set is five methods, two of them sparsities of one, and the corrected ratings span 0.003; the gap's two bootstrap constructions disagree about zero at the operational threshold—percentile $[+0.200, +0.600]$ against basic $[-0.200, +0.200]$—so what the resamples fix is its sign ($P^*(\text{gap} \leq 0) = 0.0086$, leave-one-installation-out range $[+0.200, +0.400]$) and not its magnitude. Only one pairwise comparison changes order, and it is between those two sparsities. And one candidate, Neural Cleanse, is conditional: its clean arm is exactly zero on ten of the fourteen installations, so on those its corrected rating equals its published one by construction, and dropping it returns a four-method ranking on the same thirteen installations ($\tau_a = +0.667$ for removal, gap $+0.333$, and the same percentile artifact as the headline: percentile $[+0.333, +1.000]$ against basic $[-0.333, +0.333]$, median $+0.667$, $P^*(\text{gap} \leq 0) = 0.014$). What survives is the identity with the intervention component ($\tau_a = +1.000$, $6/6$ separable pairs on that four-method set), not the size of its margin over removal. The two settings differ in more than candidate set, so we do not read them as points on one axis.

A scope condition on $A2$. Three of the eight framework defenses hold $A2$ at the policy level only. Two take a different branch on the two arms: ABL, a training-stage defense that retrains from scratch, isolates the lowest-loss examples and, on a twin with nothing to isolate, unlearns clean data instead, costing it 33 points; Neural Cleanse's detector fires on the twin arm but not the poisoned arm of seed 3, so the two arms run different computations within one installation. The third is an implementation-level use of oracle information: the framework's ANP selects its pruning threshold as the argmin of attack success on the backdoor *test* set among thresholds whose clean accuracy stays within 10% of the undefended model's (`defense/anp.py`, lines 684–687; `config/defense/anp/cifar10.yaml`, `acc_ratio: 0.1`). On a twin that objective is false-positive noise—the undefended twin scores 0.66%–0.90%—so the threshold is a test-set selection on noise under a 9-point budget, and the twin's cost runs from 0.00 (seed 10, nothing pruned) to 8.84 (seed 0, a 9.4% relative drop, just inside the budget); the information $A2$ forbids is the objective the framework optimizes, on both arms. $A2$ assumes the defense is one map applied to both arms, which holds at the policy level but not at the realised-computation level for these three. We report and flag all three (Table 30); ANP re-read under a deployer-side threshold rule that never sees attack success is Appendix S36 (registered): its tare rises from $+2.49$ to $+6.44$ on BadNets and from $+1.78$ to $+6.79$ on WaNet, because the oracle rule barely prunes a model with no backdoor. ABL is kept out of the pooled cost summary because its realised clean-arm computation collapses the twin; Neural Cleanse enters that summary only through its eight A3-successful cells, under the policy-level estimand.

S20 THE SECOND ATTACK: BLENDED ON THE SAME FRAMEWORK

We repeated the whole eleven-seed grid with BackdoorBench's blended attack in place of BadNets, eight installation seeds by the same eight defenses, both arms, everything else held fixed.

The clean arm is not blended's own twin, and we measured why. Our twin convention on BackdoorBench is to install at $\text{pratio} = 0.0002$, ten poisoned images in fifty thousand, because the framework refuses $\text{pratio} = 0$ (Section 3.1): `utils/backdoor_generate_poison_index.py` exits with "No poison sample generated" when the drawn index is empty. We left that guard in place rather than fork the release, so that the twin runs the victim's own loader and index draw and its ten images answer to A4 and to the check below; for BadNets and WaNet that is harmless. It is not harmless for blended. Evaluating three models on one attack's backdoor test set—that attack's own twin, a same-recipe model that never saw the trigger, and the poisoned model— separates the trigger's false-positive floor from what those ten images installed:

attack	poisoned images	twin's attack success	never-exposed floor	installed	A4
BadNets	10	0.0086	0.0077	0.0009	passes
WaNet	10	0.0118	0.0078	0.0040	passes
blended (3 seeds)	10	0.134	0.026	**0.108**	**fails**
blended	3	0.058	0.026	0.032	fails
blended	1	0.024	0.026	−0.002	**passes**

Whether a twin drawn at a fixed poisoning rate is clean is therefore not a property of the rate but of the attack, and blended needs a rate of its own. Three poisoned images still leave 0.058 and fail; *one* leaves 0.024 against a floor of 0.026 on seed 0—nothing—and passes; over the eight seeds it leaves 0.022–0.045, all under A4's 0.05 though not all under the seed-0 floor. That is the twin we use, so the second attack's clean arm is measured on its own pipeline rather than borrowed. We also report two alternatives as sensitivities (`blended_cells.json`): the same-seed BadNets twin, a same-recipe model that never saw this trigger, and blended's own inadmissible pratio-0.0002 twin, which pulls ANP's clean arm from $+4.54$ to $+3.89$ and I-BAU's from $+5.46$ to $+4.35$—both toward the poisoned arm, as an already-attacked control must.

Three of the framework's own defenses clear nothing here. Fine-tuning, NAD and fine-pruning pass A3 on none of the eight seeds, at mean residual attack success 0.875, 0.676 and 0.130; ANP clears $6/8$, I-BAU $5/8$, SAU $4/8$, Neural Cleanse $3/8$ and ABL $8/8$. That is the transport failure of Appendix S4 again, inside the framework rather than between rigs.

The decomposition holds; the share is again not a constant. Over the A3-passing cells, measured against blended's own admissible twin: ANP ($A2$ at the policy level only, Appendix S19) $+4.54$ $[+0.70, +8.38]$ of a reported $+7.20$ (63%), SAU $+3.94$ of $+3.33$ (118%), I-BAU $+5.46$ $[+4.26, +6.67]$ of $+3.95$ (138%), ABL $+27.8$ of $+20.3$ (137%), Neural Cleanse $+1.16$ of $+2.69$ (43%). The spread, 43% to 138% outside ABL, is the BadNets finding again on a second attack, and three of the five exceed their own reported cost.

The ordering test does not replicate, and we retract it for this cell. Only five pairs are testable here, since three defenses clear no seed, and of the four separable reversals BadNets gives, only ANP-versus-I-BAU can be tested. Its clean-arm difference keeps the BadNets sign under every clean-arm choice—-0.49 $[-5.97, +4.99]$ against the admissible twin, -3.33 $[-5.73, -0.92]$ against the borrowed BadNets twin, -0.26 $[-1.84, +1.31]$ against the inadmissible one—but separates under only the borrowed one. **We therefore claim the reordering for the BadNets cell alone.** An earlier version of this paper claimed replication here; it was resting on the borrowed column, which is exactly the substitution this appendix exists to price. More seeds would not settle it: the interval is set by ANP's clean-arm spread on this attack—the framework's test-set-selected threshold, which on a twin selects on noise (Appendix S19)—not by the five seeds.

What borrowing a clean arm across attacks costs, measured. The two clean models differ only in which attack their pipeline ran, so the substitution is the cross-attack version of the transport assumption TARE-Z makes. Paired within cell over the 26 A3-passing cells it is unbiased in the mean, $+1.50$ $[-2.81, +5.80]$, but per cell it moves the clean arm by 2.65 points on average and 8.84 at worst over the 18 cells outside ABL (6.57 and 30.23 including it). That spread cannot be attributed to the change of attack: replacing one BadNets twin by another BadNets twin, same attack and different seed, already moves ANP by 3.20 on average and 8.97 at worst (paired, Section 4.7), so what the borrowed column carries is the twin-to-twin spread on that grid, dominated by the defenses whose twin arm is unstable across seeds. Within one recipe and one attack, for the defenses whose twin arm is stable, the same substitution costs -0.02 $[-0.17, +0.12]$ (Appendix S28). Transport across attacks is therefore not separately identified here; the number bounds twin-to-twin substitution on this grid, whichever attack the borrowed twin came from.

Does the tare move with the ten images' residual attack success? A post hoc check (not registered). Pooling the seed-level tare of every admissible (attack, defense) cell of Table 4 after removing the cell's mean, and regressing it on the twin's undefended attack success over the admissible band (0.59%–4.47%; 50 twins, 49 cells, 350 seed-cells, ABL excluded), gives -0.28 points per percentage point of attack success, seed-bootstrap 95% interval $[-0.58, +0.03]$, i.e. -1.08 points across the band, an interval that covers zero; with ABL's cells the slope is -0.36 $[-1.46, +1.01]$. The sign, if any, runs the way a partial installation would (a twin scoring higher "gains" more), and its size across the band is of the order of the smallest tares in Table 4; the zero-image twins of Appendix S34 are the direct test.

S21 THE SCHEDULER CROSSOVER: THE SIGN IS SET BY ONE LINE OF THE ATTACK'S CONFIGURATION

In full. The sign is set by where the attack configuration left the victim, and we flip it on demand. `wanet/default.yaml` sets `MultiStepLR` with its first milestone at epoch 100 of a 100-epoch run, so WaNet's victim never anneals; `badnet/default.yaml` inherits the prototype's `CosineAnnealingLR`. Crossing the two schedulers across the two attacks with nothing else changed moves the clean-arm sign with the scheduler, not the attack: -1.84 $[-2.18,-1.50]$ to $+0.78$ $[+0.70,+0.87]$ on WaNet, $+0.50$ $[+0.42,+0.59]$ to -2.08 $[-2.40,-1.76]$ on BadNets. The clean arm is the distance from where the attack's run left the victim to the fixed point the defense's recovery converges to; the published cost only moves toward zero (-0.25 $[-0.66,+0.16]$) and on WaNet the poisoned arm's excess over its twin does not move (-1.17 $[-1.70,-0.64]$ to -1.03 $[-1.49,-0.58]$); on the un-annealed BadNets grid only $15/33$ poisoned-arm cells clear A3, so its residue is not read. Every patched cell in this paper is run at the benchmark's 10% poisoning rate on CIFAR-10 (5% for BPP on CIFAR-100, the rate the leaderboard reports for it there); the patch's 5% and 1% arms are Appendix S35. Input-Aware needs its own patch: its schedule has no cosine branch to fall back to, so the recommended fix is a step decay at $[30, 60, 90]$ instead of $[100, 200, 300, 400]$; run, its victim rises from 91.21 to 93.09%, confirming the same mechanism, but it still yields no admissible twin (A4 $0/5$), so the paper makes no clean-arm claim for it.

Section 4.6 reports that the clean arm reverses sign between BadNets and WaNet on BackdoorBench. This appendix reports the 2×2 crossover that turns that observation into a manipulation. **Only the `lr_scheduler` line of the attack configuration changes between a shipped cell and its swapped counterpart**: the defense recipes, the poisoning rate, the 5% clean budget, the epoch count and the learning rates are those of the grids already reported.

The configuration fact, verbatim. Both attacks inherit `config/attack/prototype/cifar10.yaml`, which sets `lr_scheduler: CosineAnnealingLR` with `epochs: 100`. `config/attack/badnet/default.yaml` overrides no scheduler and therefore inherits that cosine. `config/attack/wanet/default.yaml` overrides it with `lr_scheduler: MultiStepLR` and `steplr_milestones: [100,200,300,400]`. The attack trains for 100 epochs and WaNet's first milestone is epoch 100: **WaNet's victim model never anneals and BadNets' anneals in full.** That is the entire difference we manipulate, and it is readable in the framework's own files rather than inferred from our cells. It is not particular to WaNet: of BackdoorBench's sixteen attack configs, three (WaNet, BPP and Input-Aware) train the victim under `MultiStepLR` with a first milestone at epoch 100 of a 100-epoch run (Input-Aware builds its own from a live `schedulerC_milestones` list); the rest inherit the prototype's cosine or set their own schedule. The benchmark does not control this coordinate, and no defense paper we audit reports it.

Design and pre-registration. The pre-registration (`supplement/prereg/SCHED_PREREG.md`) was written before any of these runs existed. Two cells were added to the two shipped grids: *W-cos*, WaNet with the scheduler block deleted from its attack yaml so that it inherits the prototype's cosine, and *B-step*, BadNets with WaNet's `MultiStepLR` block added to its attack yaml. Each cell runs a poisoned arm at the framework's 10% rate and a `pratio`$=0.0002$ twin over five pre-registered installation seeds, and both arms through the three defenses that reverse sign on WaNet—fine-tuning, fine-pruning and NAD. Three predictions were recorded with the priors held at the time: **P1**, W-cos's clean arm returns to positive for all three defenses (prior 0.75); **P2**, B-step's clean arm turns negative for all three (prior 0.70); **P3**, W-cos's twin baseline accuracy rises towards BadNets' (prior 0.85). Three kill conditions were fixed with them: **K1**, if neither cell moves its sign the mechanism attribution is withdrawn and the main text keeps only the observation that the sign is unstable; **K2**, if only one cell moves, it is reported as one-directional evidence and not called a manipulation; **K3**, if annealing drops WaNet's attack success below the A0 gate of 0.90 on three or more of the five seeds, the W-cos cell is void as confounded.

Result: none of the kill conditions fired and all three predictions hold. Table 2 gives the clean-arm defense cost in each of the four cells, pooling eleven seeds per cell (the registered

five in each swapped cell and the six SCHED14 added, below); on the registered seeds alone every prediction below also holds (agreement $95/96$, WaNet-cosine $+0.82$ $[+0.72,+0.92]$, BadNets-step -2.03 $[-2.28,-1.77]$). It is an unconditional quantity—a twin carries no backdoor, so no removal-success criterion applies to it—and it is the column the crossover runs on. All 12 per-defense intervals exclude zero, and in every one the sign follows the scheduler and not the attack: annealing WaNet's victim returns fine-tuning, fine-pruning and NAD to a positive clean arm ($+0.51$ $[+0.43,+0.59]$, $+0.86$ $[+0.74,+0.97]$, $+0.99$ $[+0.86,+1.12]$), and withholding annealing from BadNets' victim turns theirs negative (-2.29 $[-2.61,-1.96]$, -1.95 $[-2.36,-1.53]$, -2.00 $[-2.30,-1.71]$). Cell by cell the agreement is $131/132$, and seed by seed (the three defenses pooled) $44/44$; the one dissenting cell is fine-tuning on BadNets+Cosine at -0.03, inside the arm whose mean is only $+0.13$ $[+0.06,+0.21]$. P3 holds too: the annealed WaNet twin rises from 0.9150 to 0.9418 and the un-annealed BadNets twin falls from 0.9371 to 0.9109. K3 did not fire—once annealed, WaNet still installs at 0.979–0.995 attack success, far above the gate. All five pre-registered seeds returned in both swapped cells; none was dropped or re-run after the fact. Six further seeds per swapped cell were then added under a separate pre-registration (SCHED14, Appendix S43) so that all four cells carry eleven; the eleven-seed pooled tares ($+0.78$ $[+0.70,+0.87]$, -2.08 $[-2.40,-1.76]$) lie inside the five-seed intervals with the same signs, all 12 per-defense intervals still exclude zero, and the cell-level agreement is $131/132$; the intervals did not narrow by the registered 30% (BadNets-step's widened), reported as P49 failed.

Scope: which column needs the admissibility machinery, and which does not. On the B-step grid the defenses frequently fail to clear the backdoor—residual attack success reaches 0.655 and only $15/33$ of the poisoned-arm cells clear at $\alpha = 0.05$—so any comparison of *reported* cost on that grid would need the A3 accounting of Appendix S3, with its conditioning and its common support, and we do not make one. The clean arm needs none of it: it is unconditional by construction, so the one column that carries the crossover is the one column in this paper that requires none of the admissibility apparatus the rest of the appendix is spent on. On the W-cos grid the same three defenses clear in $33/33$ cells.

What this establishes, and what it does not. It establishes that **the learning-rate schedule of the attack's own training run is sufficient to reverse the sign of the clean-arm cost in both directions, on this benchmark**: the same three defenses charge a never-poisoned model a positive amount or a negative one according to one line of a configuration file that has nothing to do with the defense. It does not establish that this is the only route to a sign reversal, nor that it explains WaNet's reversal in general—WaNet differs from BadNets in its trigger and its training procedure as well, we manipulated one of those differences, and a second route could produce the same reversal on a grid whose schedulers already match. What the crossover does remove is the reading that the reversal is a property of the WaNet *attack*: it is not, because the negative sign reproduces on BadNets with the attack left alone.

Table 26: The same four cells with the three columns side by side, pooled over fine-tuning, fine-pruning and NAD within a seed and then over seeds. The published cost and the residue are success-conditioned (A3); the clean arm is unconditional. On WaNet the published cost moves from -3.05 $[-3.37,-2.73]$ to -0.25 $[-0.66,+0.16]$ and the clean arm from -1.84 $[-2.18,-1.50]$ to $+0.78$ $[+0.70,+0.87]$ when only the scheduler changes, while the residue stays at -1.17 $[-1.70,-0.64]$ against -1.03 $[-1.49,-0.58]$. On the un-annealed BadNets grid only $15/33$ poisoned-arm cells clear, so its published and residue intervals are wide and we draw no conclusion from them. Printed by `supplement/analysis/build_scheduler.py`.

attack	victim's `lr_scheduler`	published cost	clean arm	removal residue (DiD)	A3 clears
WaNet	`MultiStepLR`	-3.05 $[-3.37,-2.73]$	-1.84 $[-2.18,-1.50]$	-1.17 $[-1.70,-0.64]$	$30/33$
WaNet	`CosineAnnealingLR`	-0.25 $[-0.66,+0.16]$	$+0.78$ $[+0.70,+0.87]$	-1.03 $[-1.49,-0.58]$	$33/33$
BadNets	`CosineAnnealingLR`	$+0.96$ $[+0.83,+1.09]$	$+0.50$ $[+0.42,+0.59]$	$+0.46$ $[+0.27,+0.65]$	$33/33$
BadNets	`MultiStepLR`	-2.78 $[-3.59,-1.96]$	-2.08 $[-2.40,-1.76]$	-0.74 $[-1.62,+0.13]$	$15/33$

S22 DECOMPOSING A PUBLISHED LLM DEFENSE

The two language-model defenses of Section 4.1 are our own operating points. To decompose a defense the field actually publishes, we implemented Fine-mixing (Zhang et al., 2022), whose two-stage recipe is designed for exactly this setting: mix the backdoored fine-tuned weights back toward the pre-trained weights, then fine-tune the mixture on a small clean set. We interpolate parameterwise, $\theta_{\text{mix}} = (1-\lambda)\,\theta_{\text{bd}} + \lambda\,\theta_{\text{pre}}$, and run stage two with the identical budget, optimiser and pool as our other LLM defenses (150 steps, Adafactor at 4×10^{-5}, 2000 Alpaca examples, a 3.87% clean-data budget), so the arms stay comparable. The mixing ratio is the method's own strength dial and we sweep it. Both arms of the Qwen2.5-7B pair are run at every ratio; utility is MMLU over 2000 items.

Every ratio clears the backdoor, and the decomposition bounces. Table 27 sweeps λ on a single defender draw. Residual attack success is 0.00 at every ratio, so A3 holds throughout, yet the reported cost moves between -0.45 and $+5.00$ points and the DiD term changes sign twice, ending negative at both ends of the sweep.

The reported cost is dominated by which clean examples the defender happens to draw. We replicated $\lambda = 0.5$ on both arms at three defender pool draws (Table 28). All three cleared the backdoor, but the reported cost came out -0.45, $+8.60$ and -0.70 points: a standard deviation of 5.30, which is 2.1 times the mean it is a deviation from. The removal term inherits that spread, -0.45 $[-10.86, +9.96]$, and so does the tax, $+2.93$ $[-2.18, +8.05]$. A single-run cost figure for Fine-mixing at this budget is therefore not a measurement of the defense; the clean arm is what makes that visible, because it shows the same instability on a model that never carried a backdoor. The instability is visible in training as well: three of the fourteen stage-two runs spike above a loss of 6 at a single logging step before recovering, and they are spread across both arms and two mixing ratios rather than concentrated in one cell (peak losses are a column of the supplementary cell file). We report every run; none is excluded.

Table 27: Fine-mixing (Zhang et al., 2022) on Qwen2.5-7B, one defender draw, mixing ratio swept. MMLU over 2000 items; the pre-defense poisoned model scores 0.6775 and its twin 0.6850, so the pre-defense gap is $+0.75$ points. Residual attack success is 0.00 everywhere.

mixing ratio λ	poisoned MMLU	twin MMLU	reported c_{def}	tax	removal (DiD)
0.1	0.6640	0.6520	$+1.35$	$+3.30$	-1.95
0.2	0.6590	0.6745	$+1.85$	$+1.05$	$+0.80$
0.3	0.6275	0.6575	$+5.00$	$+2.75$	$+2.25$
0.5	0.6820	0.6480	-0.45	$+3.70$	-4.15
0.7	0.6560	0.6520	$+2.15$	$+3.30$	-1.15

Table 28: Fine-mixing at $\lambda = 0.5$, both arms, three independent defender pool draws from the same Alpaca corpus at the same 3.87% budget. All six runs cleared the backdoor. Intervals are 95% t-intervals over the three draws.

	draw 2	draw 3	draw 4	mean [95%]
reported c_{def} (pt)	-0.45	$+8.60$	-0.70	$+2.48$ [−10.68,+15.65]
intervention tax (pt)	$+3.70$	$+4.50$	$+0.60$	$+2.93$ [−2.18,+8.05]
interaction, DiD (pt)	-4.15	$+4.10$	-1.30	-0.45 [−10.86,+9.96]

S23 EXTENSION TO MACHINE UNLEARNING

Machine unlearning has the analogous baseline-accounting problem, and there the reference backdoor benchmarks omit is already standard. On TOFU (Maini et al., 2024) the clean twin is *exact*, a from-scratch retrain on the retain split, and it removes the forget set at $+0.82$ points of utility ($[-1.17, +2.81]$, five seeds), similar across forget-set sizes (1–10%; Table 24), while the utility drops that published TOFU reproductions report for algorithms scored against that same gold span -2 to 62 points (Fan et al., 2025; Mekala et al., 2025): a cross-study span read off those papers, not a quantity we recompute.

S24 PER-DEFENSE RE-RANKING

Per-defense detail behind the re-ranking of Section 4.5 (the slope diagram is Figure 5(b)).

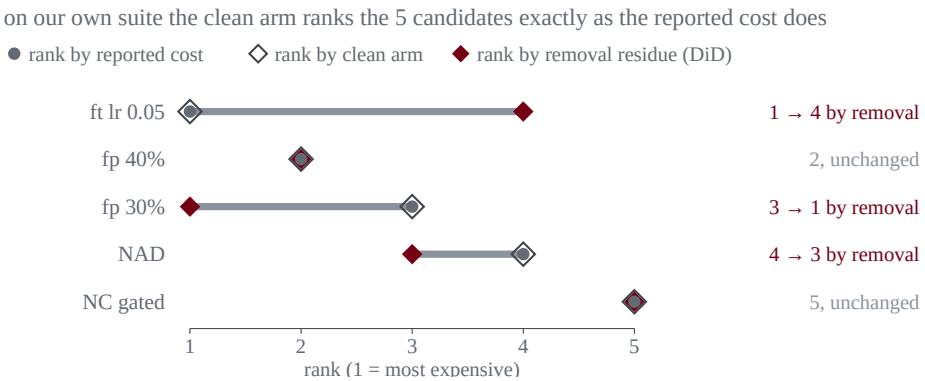


Figure 7: **On our own suite the clean arm ranks the candidates exactly as the reported cost does ($\tau_a = +1.000$, $10/10$ pairs separable), which is the identity of Eq. 5 made visible; only the removal residue moves anything.** The five candidate defenses on CIFAR-10/BadNets over the thirteen installations of the common support, where every candidate clears the backdoor and keeps an admissible twin. Grey dot: rank by reported cost. Hollow diamond: rank by the clean arm—it coincides with the grey dot on every row. Crimson diamond: rank by the DiD removal residue. **The rows are drawn in order of cost magnitude, not of the DER rating every τ_a we quote is computed on**: on the drawn rule the published-versus-removal agreement is $+0.200$ with a gap of $+0.800$ and three of the five rows moving, while on the DER rule of Table 12 it is $+0.800$ with a gap of $+0.200$ (percentile $[+0.200, +0.600]$, basic $[-0.200, +0.200]$—the two constructions disagree about zero, Appendix S3), $3/10$ pairs separable and one adjacent transposition. The corrected ratings all lie within 0.003 of one another, so the corrected rank intervals are wide and we report the ordering as a consequence of the identity rather than a result; the separable reversals that are a result are on BackdoorBench (§4.6). NAD enters this suite only in this version: its teacher had been built at a learning rate that leaves the backdoor intact, and at a teacher that loses it NAD clears $13/14$ (Appendix S30). All rank correlations are Kendall τ_a on unrounded scores; rank intervals are conditional on the selected candidate set and the realized common support, both adopted during internal review and reported with the alternatives they replaced (Table 11), and that selection uncertainty is not propagated by the bootstrap. The four methods excluded by the candidate gate are tabulated in Appendix S3 and below.

Table 29: **On the community's own benchmark the decomposition replicates and the ordering result does not.** BackdoorBench's shipped attack+defense framework, BadNets on PreAct-ResNet-18 (the framework's default), eleven independent installation seeds under the protocol of Appendix S19, frozen before the eight confirmatory seeds. Cost rows average the seven defenses other than ABL, a training-stage defense retrained from scratch, whose clean-twin drop exceeds 30 points, dropping A3-failed cells first as elsewhere: Neural Cleanse contributes eight of its eleven cells, the rest all eleven. The ABL row and the 21 pairwise comparisons include ABL. The ranking rows use different sets: the candidate set is the seven methods that pass the method-level gate, which excludes Neural Cleanse and includes ABL, and its alignment gap is $+0.190$ $[-0.190, +0.476]$; the rank-correlation row below is the prespecified without-ABL sensitivity over six methods; the pairwise comparisons run over the seven gate-passing candidates. The entries close at the point-estimate level as Eq. 5 requires: $+2.42 - 2.06 = +0.36$ (paired minus gap) and $+0.36 + 1.98 = +2.34$ (removal plus clean arm). WaNet is a third attack on the same framework under the same frozen protocol, eleven installation seeds: three fail A0 (attack success 0.37–0.75 against 0.94–0.99 elsewhere), leaving a success-conditioned common support of five seeds over the six candidate defenses (pre-defense gap $+1.43$ $[+0.48, +2.39]$). Its ranking gap is $+0.000$, bootstrap $[-0.667, +0.400]$ with $P^*(\text{gap} \leq 0) = 0.7842$, and the exact sign-flip enumeration separates $0/15$ pairs in *both* accountings, so that zero is an unpowered grid rather than a null. On the four confirmatory seeds the corrected order is *identical* to the published one ($\tau_a = +1.000$) but the intervention ordering is not ($+0.600$), so the gap there is -0.400 $[-0.667, +0.133]$ with $P^*(\text{gap} \leq 0) = 0.9382$: on those seeds it runs the other way. The powered test on this grid is the pairwise envelope (Appendix S19). †ANP holds $A2$ at the policy level only: the framework selects its threshold by test-set attack success (Appendix S19).

	published accounting	twin-corrected
clean-accuracy cost term (pt, 11 seeds) ↓	$+2.34$ [+2.10,+2.57]	$+0.36$ [−0.26,+0.98] (DiD)
pre-defense gap, twin − poisoned (pt)	$+2.06$ [+1.90,+2.23]	
paired post-defense contrast (pt)	—	$+2.42$ [+1.81,+3.04]
clean-arm defense cost (pt)	—	$+1.98$ [+1.50,+2.46]
with ABL, eight defenses (pt)	$+3.25$ [+3.01,+3.49]	-2.79 [−3.69,−1.88] (DiD)
attack-success drop (%)	93–95 (near-constant)	
SD across the six non-ABL defenses (pt)	2.40 (poisoned arm)	2.08 (backdoor-free twin)
rank correlation (without-ABL, 6)	—	$+0.733$ rem., $+0.200$ interv.
separable pairs (BH, q=0.05)	$20/21$	$18/21$
third attack (WaNet): mean, n=5 cond.	-1.32 [−2.18,−0.46]	-0.63 [−1.94,+0.68] (DiD)
clean-arm defense cost, n=5 cond.	—	-0.69 [−1.46,+0.09]
clean-arm defense cost, n=11 uncond.	—	-0.61 [−0.93,−0.29]
paired post-defense contrast, n=11 (Table 29's convention)	—	$+1.00$ [+0.55,+1.45]
fine-tuning alone, paired, n=11, BadNets / WaNet	$+2.94$ [+2.75,+3.14]	$+0.34$ [+0.18,+0.50]
clean arms of reversed sign, n=11	$3/8$ excl. zero: ft -1.99, fp -1.68, NAD -1.85	
negative reported cost, n=5 cond.	$5/6$ candidates, every interval excl. zero	
largest (ANP†), n=5 cond.	$+4.37$ [+0.46,+8.28]	$+2.28$ [−2.07,+6.64]
seeds installing below the A0 gate	$3/11$ (success 0.37–0.75)	

Table 30: **At this attack the non-removal share is a property of the defense, not a constant; on WaNet the same eight do not even agree on its sign (Table 29), so it is a property of the cell.** Per-defense decomposition on BackdoorBench over the eleven installation seeds, 95% t-intervals. Every defense levies a clean-arm cost whose nominal interval excludes zero ($6/8$ at $0.05/8$, $8/8$ under BH $q = 0.05$; excluding ABL, $5/7$ at $0.05/7$ and $7/7$ under BH, the main text's convention). Four of the seven unconditional methods have non-positive interaction (DiD) point estimates: the identical defense costs the never-poisoned twin more than it costs the poisoned model, which is what a rating that charges the whole drop to removal cannot represent; in paired form (A0 seeds, Table 4) only ABL is non-positive (fine-pruning $+0.80$, SAU $+0.76$, I-BAU $+0.21$), the two forms differing by $g = +2.06$, and which part of either is removal is not identified (Section 3.1). Every row averages that defense's A3-passing cells, so Neural Cleanse's row is over its eight successes rather than all eleven seeds; an earlier version printed its row over all eleven, the only row in the table that used a different convention. Neural Cleanse is listed for completeness; the framework's detector does not fire on three of its eleven seeds, so the candidate gate excludes it there at a mean residual attack success of 0.267. In our own harness, with the trigger reverse-engineered in pixel space, the same method fires on all fourteen and clears every one (Appendix S26); the two rigs disagree about this defense as they do about several others.

Defense	reported c_{def} (pt)	clean arm (pt)	interaction, DiD (pt)	+noise	+shortage
fine-tuning	$+1.01$ [+0.84,+1.19]	$+0.13$ [+0.06,+0.21]	$+0.88$ [+0.68,+1.08]	$+2.02$	$+0.47$
NAD	$+2.10$ [+1.89,+2.30]	$+0.33$ [+0.24,+0.42]	$+1.77$ [+1.52,+2.02]	$+2.90$	$+1.35$
ANP†	$+6.50$ [+5.73,+7.27]	$+2.49$ [+0.39,+4.60]	$+4.01$ [+1.69,+6.32]	$+5.14$	$+3.59$
fine-pruning	-0.22 [−0.36,−0.08]	$+1.05$ [+0.84,+1.25]	-1.27 [−1.60,−0.93]	-0.13	-1.68
SAU	$+1.39$ [+1.11,+1.67]	$+2.70$ [+2.31,+3.08]	-1.31 [−1.80,−0.82]	-0.17	-1.72
I-BAU	$+3.85$ [+3.05,+4.66]	$+5.70$ [+3.99,+7.42]	-1.85 [−3.93,+0.23]	-0.71	-2.26
ABL†	$+9.39$ [+8.12,+10.7]	$+33.2$ [+29.1,+37.3]	-23.8 [−28.3,−19.3]	-22.7	-24.2
Neural Cleanse†‡	$+1.53$ [+0.99,+2.07]	$+1.31$ [+0.10,+2.51]	$+0.22$ [−0.85,+1.29]	$+1.36$	-0.19
pooled, 7 excl. ABL	share of c_{def} left for the interaction		15%	64%	-3%

The last two columns apply the headroom correction of Appendix S5 to every row or to none, once for each degradation route: an arm degraded by label noise implies -1.14 at this benchmark's gap, an arm degraded by training on less data $+0.41$. The routes disagree in sign, and the poisoned arm's own deficit resists these defenses the way the second route's does, so we apply neither and report the uncorrected column in bold as the measurement. The clean arm needs no correction. $^{\dagger}A2$ at the policy level only: ABL and Neural Cleanse run different computations on the two arms, and ANP selects its pruning threshold by test-set attack success under a 10% accuracy budget, on a twin a selection on false-positive noise (Appendix S19). ABL is excluded from the pooled row; Neural Cleanse enters it through its eight A3-successful cells. ‡conditional; excluded from the ranking by the method-level gate.

Table 31: Sensitivity of the field's ordering to how its cost term is defined. Nine defenses, CIFAR-10 / BadNets, 5% clean budget, all fourteen independent installations, corrected defender draw. The last column gives, within the five candidate rows, the rank under the published, intervention-component and DiD-removal DER ($\tau_a = +0.800$ and $+0.200$); primary inference uses the thirteen-installation common support, not these fourteen, which carry the rank column only. ¶ANP is excluded on A3, not on A4: under its own accuracy budget it prunes too little to clear this attack (Table 1, Appendix S3). A sensitivity analysis, not a proposed score. ‡Neural Cleanse is conditional and gated as its authors specify; its clean arm no-ops in ten of the fourteen installations, so on those its two columns coincide by construction, and on the other four it charges a never-poisoned model 2.75–3.67 points (Appendix S26). Over all nine rows the published-versus-corrected agreement is $\tau_a = +0.778$. These are the 14-installation values under the candidate gate; the primary common-support analysis, over the five gate-passing methods on the thirteen installations they all clear, is Table 12. NAD and fine-pruning at 30% agree to five decimals on the published score (0.9634000 against 0.9633964, a margin of 3.6×10^{-6}) and are ordered on the unrounded value. §marks methods whose *mean* residual attack success exceeds the 5% method-level gate. They are not universally inadmissible—A3 is an installation-level event and some of their cells clear it (Appendix S3)—but they are excluded from the candidate set for cross-method ranking and enter no primary figure.

Defense	A after	DER published	DER corrected	rank pub./interv./DiD
Neural Cleanse	0.008	0.9741	0.9754	1/1/1
NAD	0.022	0.96340	0.9705	2/3/5
fine-prune 30%	0.016	0.96340	0.9731	3/2/3
fine-prune 40%	0.014	0.9582	0.9745	4/4/2
fine-tune lr 0.05	0.018	0.9520	0.9725	5/5/4
fine-prune 20%§	0.062	0.9454	0.9507	—
ANP§	0.509	0.7183	0.7250	—
fine-tune lr 0.02§	0.613	0.6698	0.6746	—
fine-tune lr 0.01§	0.856	0.5522	0.5535	—

S25 INDEPENDENTLY RE-TRAINED LANGUAGE-MODEL SEEDS

Table 32 gives the per-defense detail behind the pooled language-model figures of Section 4. Each seed is a full end-to-end re-run: poisoning, clean twin, and both defenses retrained from the base model.

Table 32: Language-model costs under eight independently re-trained installation seeds (Qwen2.5-7B-Instruct, MMLU over 2000 items; each seed is a full re-run of poisoning, twin, and both defenses with the corrected, independent defender draw; installation-level A3 successes are $7/8$ for each defense, fine-tuning missing it in installation 15 at residual ASR 0.09 and fine-pruning in installation 17 at 0.25; restricting to the six installations where both succeed gives a complete-case removal aggregate of $+1.60$ $[+0.79, +2.42]$ against a reported cost of $+12.05$ on those same six, so the headline does not depend on variable-composition averaging). Per-defense intervals are 95% over installation seeds; the aggregate averages the admissible defenses within each seed ($\mathrm{df} = 7$). Pre-defense gap $g = +0.04$ $[-0.19, +0.27]$.

Defense	c_{def} (pt)	c_{int} (pt)	$c_{\mathrm{int}}^{\mathrm{DiD}}$ (pt)
full fine-tune lr 4×10^{-5}	$+8.74$ [+5.52,+11.97]	$+2.78$ [−0.19,+5.75]	$+2.77$ [−0.17,+5.71]
fine-prune 20%	$+14.68$ [+11.41,+17.95]	$+0.25$ [−1.94,+2.44]	$+0.17$ [−2.11,+2.46]
install level (n=8)	$+11.46$ [+8.94,+13.97]	$+1.42$ [+0.69,+2.14]	$+1.37$ [+0.52,+2.23]

S26 CONDITIONAL DEFENSES

Policy-level consistency (A2). A conditional defense is one function at the policy level—detection, then repair only if the detector fires—so the twin arm is well-defined and simply takes the no-op branch on a clean model. With the trigger reverse-engineered in pixel space (Appendix S12) Neural Cleanse (Wang et al., 2019) fires on all fourteen poisoned models, names the true target class in each, clears the backdoor in all of them (mean residual attack success 0.008), and charges $+0.78$ $[+0.60, +0.96]$ points. Its gated twin arm no-ops in ten of the fourteen installations, so on those the corrected and reported costs coincide by construction.

The gate does not make the repair cheap; it makes it rare. That reading is testable, and we tested it: forcing the repair branch on every never-poisoned twin costs $+3.40$ $[+3.21, +3.59]$ points, more than four times what the same defense charges the poisoned models it does fire on. The price is a property of the repair, not of our patching schedule: at Wang et al.'s own recipe, one epoch over 10% of the training set with a fifth of each batch patched, the forced repair costs $+3.51$ $[+2.79, +4.23]$. The asymmetry has a mechanism—on a poisoned model the search recovers a small, real trigger and unlearning it is cheap, while on a clean model it must invent a large diffuse pattern and unlearning that damages the network—and it is invisible without the twin, because the gated clean arm reads exactly zero. The detector also false-positives on four of the fourteen twins, naming a wrong class each time at an anomaly index just over the threshold of 2, and on those it charges a never-poisoned model 2.8–3.7 points. **The confound we identify is a property of unconditional defenses and of false-positive repairs**; for a gated policy it is priced by the false-positive rate times the forced-repair cost, and a benchmark scoring gated and ungated defenses under one rating should say which it is scoring.

S27 THE SURROGATE'S TWO TIERS

The estimator has two tiers, and we report both. **Within the tested architecture and recipe** its error is a clean-versus-clean difference: on Qwen2.5-7B, four independently trained clean models used one at a time give DiD-form surrogate-minus-oracle errors from -0.5 to $+3.0$ points, mean $+1.7$, against a twin-arm cost of 8.9 (Table 33); averaging them brings the error to the mean, a fifth of the clean-arm cost it replaces. On vision the instability is concentrated in ANP (Table 35). **Across architectures** it is qualitative: Llama-3.2-3B as a surrogate for Qwen2.5-7B (different family, 2.3× scale gap) carries a -1.29-point systematic bias (range -2.47 to -0.42 over five fixed defense configurations, all same-signed; Table 34; we report no interval because the configurations are design

points, not replicates), which supports the qualitative verdict that the reported cost is overstated, but not a quantitative decomposition. The boundary is set by a third architecture—Mistral-7B catastrophically forgets under the same recipe (a clean Mistral loses 26.0 MMLU points to the identical defense, lr 4×10^{-5} for 150 steps, against 7.7 for a clean Qwen; both drops are within-model, each arm scored on its own 2,000-item pool, `supplement/data/llm_arm_evaluations.json`), an instance of the path-dependence formalized by Yu et al. (2025)—and a same-recipe degradation probe identifies that catastrophic case without an oracle. The probe does not by itself certify exchangeability or guarantee surrogate validity; it screens out the regime in which the surrogate is known to fail.

Deployment scenario. A model provider receives a suspected or externally supplied checkpoint and can test whether a defense suppresses a known or independently discovered trigger, but does not hold the poisoned training examples needed to reconstruct a matched twin. It does hold a clean model of the same architecture and training recipe—a prior production replica or an independently trained release candidate. TARE-Z uses that model to estimate the clean-arm correction at the cost of one defense run; the degradation probe screens out the catastrophic regime observed here, and the removal-success evidence must still be obtained independently.

Table 33: Same-recipe surrogates on Qwen2.5-7B-Instruct: continued fine-tuning at lr 4×10^{-5} for 250 steps applied to the paired twin and to four independently trained clean models (MMLU over 1,000 items, in points). The DiD-form surrogate error is the twin's drop minus the surrogate's drop; it does not involve the poisoned arm, so these rows are unaffected by the defender-draw correction of Appendix S30. The poisoned-arm reference row an earlier version carried was affected and is omitted.

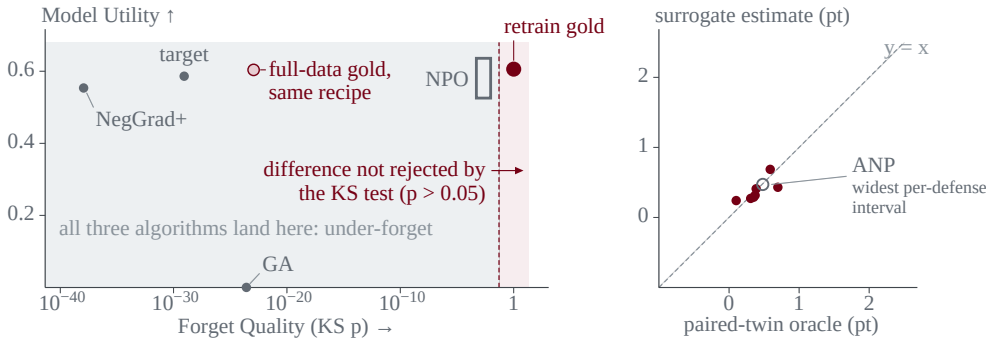
clean model	$Q(\theta)$	$Q(\mathcal{D}_b\theta)$	drop	error vs. twin
paired twin (oracle)	70.7	61.8	8.9	—
clean seed 1	71.2	65.3	5.9	+3.0
clean seed 2	70.4	64.0	6.4	+2.5
clean seed 3	70.3	63.1	7.2	+1.7
clean seed 4	70.9	61.5	9.4	−0.5
mean of the four	70.7	63.5	7.2	+1.7

Table 34: Cross-architecture surrogate: Llama-3.2-3B's clean-arm defense drop substituted for Qwen2.5-7B's, per defense configuration (points; every value, both arms and both models, is MMLU over 2,000 items; n is the number of defense seeds averaged). The error, Qwen drop minus Llama drop, is the error of the DiD-form surrogate estimate: mean -1.29, range -2.47 to -0.42, all five same-signed; the configurations are fixed design points, so no interval is attached.

configuration	Qwen twin drop (n)	Llama twin drop (n)	error
fine-tune, lr 3×10^{-5}	5.22 (3)	6.60 (1)	−1.38
fine-tune, lr 4×10^{-5}	7.20 (3)	8.80 (3)	−1.60
fine-tune, lr 5×10^{-5}	9.83 (3)	12.30 (3)	−2.47
fine-prune 20%	12.40 (2)	13.00 (3)	−0.60
fine-prune 40%	23.65 (2)	24.07 (3)	−0.42

S28 SURROGATE CALIBRATION, PER DEFENSE

Table 35 gives the per-defense detail behind the pooled figures of Section 4.7. On these four installations the per-defense oracle values outside ANP span only 0.6 points, so an ordering comparison is uninformative there (Kendall $\tau_a = +0.857$ over all eight, $+0.905$ without ANP); the calibration claim is the error column, not an ordering.



(a) on TOFU only the control reaches p > 0.05

(b) a same-recipe clean model tracks the twin correction

Figure 8: **Two checks on the control arm. (a)** On TOFU the from-scratch retrain (crimson) is the only point outside the shaded under-forgetting region: it forgets perfectly by construction—it *is* the reference the KS test compares against, so its $p = 1$ is a definition, not a finding—at a utility that does not resolve from the full-data model's, while none of the three algorithms we run reaches that forget quality, and harder NPO settings lower utility without reaching it. GA collapses entirely; NPO is drawn as a box over seeds and harder settings; the open marker is the full-data gold under the same recipe, which has forgotten nothing and lands with the target at $p \approx 10^{-23}$, the anchor at the far end of the axis. Every p drawn here is recomputed under one definition by `analysis/tofu_fq.py`, which reproduces the three values the eval dumps carried (one seed shown; five-seed means in Table 24). **(b)** Pairing is not essential for the mean correction in this suite: an independently trained clean model of the same recipe (Eq. 6) puts all eight unconditional defenses within 0.3 pt of $y = x$ in the mean (pooled error -0.02 $[-0.17, +0.12]$ pt, largest single defense-by-installation error 0.88 pt). The open marker is ANP, whose *oracle* estimate is the most installation-unstable of the eight, so its wider per-defense interval cannot be attributed cleanly to either side—it marks the boundary of the estimator's precision, not an outlier we discard. CIFAR-10, eight defenses over four installations; each point is a defense's mean over the four.

Table 35: The surrogate agrees with the oracle on average and tracks all eight unconditional defenses closely. CIFAR-10, 5% clean budget, four installations; surrogates are independently trained clean models with no pairing to the poisoned run. Neural Cleanse is excluded because its gated clean arm usually no-ops, leaving no clean-arm intervention correction for the surrogate to estimate. The last column is what a practitioner using the surrogate instead of the true twin would be wrong by; per-defense rows give the paired form, and the four pooled rows give both the paired and the DiD form. All intervals are t-intervals over the four installations, pooled rows over the four installation means. Two independently trained clean models serve as surrogates for each installation and their estimates are averaged; both arms use the identical defense implementation, including the BN-masked fine-pruning. Fine-tuning at lr 0.01 and 0.02 do not clear the backdoor on these installations (NAD, on its corrected teacher, clears three of four), so their oracle column is a post-defense contrast rather than an identified removal cost; the surrogate's error is a clean-versus-clean quantity either way. ANP no longer needs a separate reading: with its reference mask semantics and accuracy-guarded threshold its surrogate error is -0.01 $[-1.02, +0.99]$ like the rest, and the pooled error is the same with and without it. An earlier implementation of ANP made it the one outlier and dominated the pooled row (Appendix S30).

Defense	oracle c_{int} (pt)	surrogate $\hat{c}_{\text{int}}$ (pt)	difference (pt)
fine-tune lr 0.01	$+0.39$	$+0.41$	$+0.02$ $[-0.38,+0.42]$
fine-tune lr 0.02	$+0.59$	$+0.69$	$+0.10$ $[-0.14,+0.34]$
fine-tune lr 0.05	$+0.70$	$+0.43$	-0.27 $[-0.65,+0.12]$
fine-prune 20%	$+0.10$	$+0.24$	$+0.14$ $[-0.05,+0.32]$
fine-prune 30%	$+0.37$	$+0.32$	-0.06 $[-0.20,+0.09]$
fine-prune 40%	$+0.31$	$+0.27$	-0.04 $[-0.35,+0.27]$
NAD	$+0.35$	$+0.29$	-0.06 $[-0.57,+0.45]$
ANP	$+0.48$	$+0.47$	-0.01 $[-1.02,+0.99]$
pooled paired, all eight	—	—	-0.02 $[-0.17,+0.12]$
pooled paired, excl. ANP	—	—	-0.02 $[-0.10,+0.05]$
pooled DiD, all eight	—	—	-0.03 $[-0.48,+0.42]$
pooled DiD, excl. ANP	—	—	-0.03 $[-0.52,+0.46]$

S29 THREE COST ESTIMATORS, SIDE BY SIDE

In full. What the twin attributes to the interaction is estimator- and defense-dependent under either form. Table 25 gives every setting under both, and matched denominators put none above 55% (Appendix S8); on CIFAR-10 a reported $+2.51$ leaves $+0.02$ $[-0.31, +0.34]$ by difference-in-differences and $+0.43$ $[+0.30, +0.55]$ paired. Over the 8 CIFAR installations the paired estimate's slope in g is -0.16 $[-0.46,+0.10]$ (exact permutation $p = 0.3950$ over 40320 orderings); since $c_{\text{int}}^{\text{DiD}} = c_{\text{int}} - g$ cell by cell, the DiD slope is that minus one, -1.16 $[-1.46,-0.90]$, an identity rather than an independent finding. Qwen's $+1.37$ of $+11.46$ is the only all-installation estimate above zero (BackdoorBench's confirmatory seeds alone give $+0.52$ $[+0.06,+0.99]$) and the one we hold most lightly; its operating points are expensive because nothing cheaper works there (Appendix S12).

Table 23 reports all three estimators of Section 3 on the CIFAR defense families over eight independent installations. The difference-in-differences correction subtracts the pre-defense gap ($+0.41$ pt, CI $[+0.15, +0.68]$ over the eight installations), so it shifts every c_{int} down by roughly that amount; on this grid that is the difference between a paired contrast that resolves ($+0.43$ $[+0.30, +0.55]$) and a DiD that does not ($+0.02$ $[-0.31, +0.34]$); Appendix S5 reports the slope of each estimator in g over the eight installations—the paired estimate's is -0.16 $[-0.46,+0.10]$, and the DiD's is that minus one by identity, -1.16 $[-1.46,-0.90]$. The qualitative picture is unchanged under all three: the reported cost c_{def} grows with defense strength while the removal cost does not.

On the language model the two estimators nearly coincide on the eight-installation data ($c_{\text{int}} = +1.42$ and $c_{\text{int}}^{\text{DiD}} = +1.37$ pt at the installation level; pre-defense gap $+0.04$ pt). No conclusion in the paper depends on which estimator is used; Table 25 prints both, and Section 3.1 states why neither is corrected into a removal cost.

An attempt to manipulate g experimentally, reported as inconclusive. Because the two estimators differ by g, the natural design is to vary g at the source. We installed BadNets at 1% and 10% poisoning instead of the headline 5%, three installations each, and ran four defenses on each (supplementary `cifar_poisonrate_cells.csv`). The manipulation did not move the manipulated variable: g is $+0.94$ $[-0.22, +2.10]$ at 1% and $+1.15$ $[-0.29, +2.58]$ at 10%, against $+0.53$ $[-0.18, +1.24]$ at 5% on the same seeds, three overlapping intervals; pre-defense attack success is 0.94 and 0.97. What the poisoning rate does change is removal success, which falls at the low rate: at 1% only fine-pruning at 40% clears the backdoor in all three installations, and fine-tuning at lr 0.02 in none. A second reason not to lean on this design: the 1% and 10% arms were installed for 30 epochs where the headline installations run 40, so their pre-defense gaps mix the poisoning rate with a training-budget difference and are not comparable to the headline g on equal terms. We therefore report the slope in g of Appendix S5 from the variation the installations supply on their own (g from -0.06 to $+0.83$), and record this design as inconclusive rather than as evidence.

S30 SELF-CORRECTION DETAILS

In full. Whether a named defense removes the backdoor is a property of the cell, not the method. At BackdoorBench's own operating point its fine-tuning leaves a mean residual attack success of 0.015 on its installations and 0.369 on ours, and on a second attack inside that framework its own fine-tuning, NAD and fine-pruning clear none of eight seeds. One such gap was ours, not the installation's: our NAD cleared $0/14$ until we rebuilt its teacher, which we had been fine-tuning at a rate that leaves the backdoor intact; at a teacher that loses it NAD clears $13/14$, and its clean arm of $+1.64$ exceeds its entire reported cost of $+1.49$. That a defense's verdict depends on how intensely the model was trained on the poison is established (Zhu et al., 2025; Min et al., 2023; Zhu et al., 2023a); we cite it as the explanation rather than offer the failure as a finding (Appendices S4, S20, S30).

The defender's clean data was not independent of the attacker's poison draw. In the vision harness the poisoned indices and the defender's clean subset were drawn from the same seeded stream with the same call, so for every installation the two index sets were bit-identical: the defender fine-tuned on exactly the images whose labels the attacker had flipped, now correctly labelled. The language-model harness had the same defect, the defender's pool being the installation pool. The bias runs in the direction of our own thesis, because an oracle repair set makes removal on the

poisoned arm unusually cheap and unusually reliable. We found it when a reviewer read the shipped code, fixed the draw to an independent stream, and re-ran every affected cell: the entire CIFAR grid, the eight Qwen installations, which had to be rebuilt from scratch because their weights had been deleted, the Llama supporting row, the ranking grid and the Fine-mixing decomposition. Three results moved materially. Success rates fell, from $47/48$ to $37/48$ on the CIFAR headline grid and from $18/32$ to $0/32$ for last-convolution fine-pruning, which under a realistic defender never reaches the 5% residual-attack threshold at any sparsity we tried. The CIFAR removal estimate fell from $+1.05$ $[+0.78, +1.33]$ to $+0.02$ $[-0.31, +0.34]$. The Qwen estimate rose and changed sign, from -0.24 $[-1.45, +0.98]$ to $+1.37$ $[+0.52, +2.23]$. Everything reported here is post-fix; the pre-fix numbers appear only in this paragraph. BackdoorBench and TOFU were unaffected, since they use their own data pipelines.

Three defects in our own fine-pruning, found by re-reading the reference method. Fine-pruning ranks channels by activation and then recovers by fine-tuning. Our first implementation zeroed the selected channels once and then ran the recovery loop, so gradient descent refilled them: by the end of recovery the "pruned" network was a fine-tuned dense one. We now re-apply the channel mask after every optimizer step. Second, the last-convolution scope selected `layer4.1.conv2`, whose output the residual identity path routes around, so masking it changed almost nothing; the scope now ranks and masks the post-residual `layer4` output the classifier actually sees, and at 95% sparsity 26 of 512 output channels remain non-zero. Third, zeroing a convolution's weights does not silence its channel, because the following BatchNorm adds β: on the masked channels the post-norm output had mean 0.331, not zero. We now zero the matching BatchNorm γ and β as well, and all 153 masked channels measure silent. The three fixes together move the CIFAR headline from $+2.81$ to $+2.51$ points reported and leave the removal term at $+0.02$ $[-0.31, +0.34]$ against $+0.04$ before (Table 8); they do not move the paper's conclusion, and we report them because the first version would have let a reader reproduce a number our own description did not describe.

A bookkeeping error the rebuild caught. The framework does not run the defended clean arm in any of its shipped attack–defense pipelines, so every difference-in-differences here is assembled from its per-cell summaries by our own scripts. An earlier internal version of this paper read one such assembled quantity as a paired post-defense contrast and subtracted the pre-defense gap a second time, reporting a removal component of -2.23 and a clean-arm cost of 4.48 points where the correct three-seed values were -0.09 and $+2.34$. Every BackdoorBench quantity here is rebuilt from the framework's own per-cell summaries; the eleven-seed values are $+0.36$ $[-0.26, +0.98]$ for removal and $+1.98$ $[+1.50, +2.46]$ for the clean arm (Table 29). The error was ours, not the framework's, and it inflated the correction in our favour, which is why we state it here rather than in a changelog.

Clean-data budget is the controlling variable. Holding the model (ResNet-18), the poisoning run, and the defense (fine-pruning at 30%) fixed, and varying only the fraction of clean data available to the defender, the reported cost moves monotonically to zero while the removal term stays within half a point of it; the full seven-point sweep is Table 22, Appendix S16. The same parameter count spans four accuracy points as the budget alone changes, which is why we treat the defender's clean-data budget as a protocol coordinate and report all vision results at the field-standard 5%.

Clean–clean variability calibration. The placebo arm pairs two never-poisoned *vision* models differing only in seed and applies the identical defense to both. Over four such pairs the difference in defense cost is $+0.06$ pt with a 95% t-half-width of 0.74 pt (an earlier version quoted 0.39, which was $1.96\times$ the standard error rather than the t interval this paper otherwise uses). The band bounds the *vision* claim only: the DiD estimate, $+0.02$ $[-0.31, +0.34]$, sits inside it, and so does the paired contrast, $+0.43$ $[+0.30, +0.55]$, which resolves by its own interval. It does *not* transfer to the language model, which is a different instrument; an earlier version of this paper applied it there, and Appendix S30 records the correction. The language-model claim is stated at the installation level (n=8, $\mathrm{df} = 7$): a $+1.37$ pt removal cost with 95% CI $[+0.52, +2.23]$ (Table 32). A direct language-model placebo—five independently trained clean models of the same recipe, each put through the identical defense—pays 5.9–9.4 pt, and the ten pairwise differences among those five costs have one standard deviation of ± 2.1 pt. Repeating the construction on the eight headline installations' own twin arms gives ± 3.2 pt; we report the smaller ± 2.1 descriptive scale, whose upper edge the removal interval just crosses ($+2.23$ against $+2.10$). We draw no equivalence conclusion from the

band: the language-model claim in Table 7 is the interval itself, whose upper end $+2.23$ is 19% of the reported 11.46.

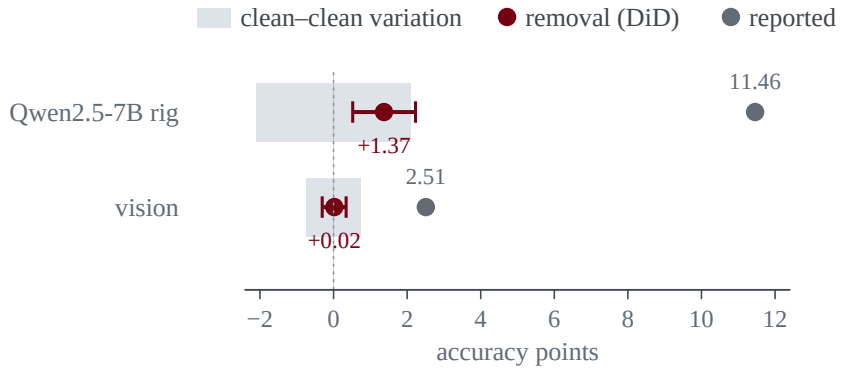


Figure 9: **Both reported costs sit far outside the clean-versus-clean variation of their own rig; both interaction estimates sit near it.** DiD estimates in accuracy points with 95% intervals over installations (n=8 in both settings). The shaded band is clean-versus-clean variation, shown only as a descriptive, rig-specific reference: ± 0.74 pt on vision (the 95% t-half-width of four clean–clean pair differences) and ± 2.1 pt on the language model (one standard deviation of five clean models' pairwise cost differences). These are different statistics on different instruments and are not compared across rigs. The statistical conclusions rest on the intervals themselves: the Qwen interval excludes zero and its upper endpoint, $+2.23$ points, is 19% of the reported-cost point estimate, while the vision interval includes zero.

The threshold metric hides no margin. `asr50` is thresholded, so it could saturate while an unthresholded margin kept growing. It does not: over a $20\times$ dose sweep the unbounded trigger-vs-clean KL divergence slightly *falls* rather than rises, and thresholds $50/90/99$ agree throughout (Appendix S13).

A conditional defense, first run unconditionally, then given a trigger it could reach. Neural Cleanse repairs only when its anomaly index exceeds 2; our first implementation repaired regardless. We caught that only because a reader asked why a defense would damage a model it had declared clean. Gating it exposed a second defect. The trigger search composed its candidate pattern on the loader's already-normalised tensors and squashed it through $\tanh$, so the pattern could not exceed 1.0 in those units while the BadNets white patch sits at $(1 - 0.4914)/0.2470 = 2.06$; the optimiser compensated with a larger mask, which inflated every ℓ_1 norm and flattened the outlier score the detector reads. The weight on that norm was also a fixed constant rather than the published dynamic schedule. Reverse-engineering in $[0, 1]$ pixel space with that schedule, the detector fires on $14/14$ poisoned models against $9/14$ before, names the true target class every time, and clears the backdoor everywhere (mean residual attack success 0.008 against 0.368) at $+0.78$ points against $+4.24$ over the cells that cleared before. It becomes a ranking candidate, bringing the candidate set to five, and it is from a family that is not a strength variant of any other member. The scope condition the gate exposes—the confound is a property of unconditional defenses and of false-positive repairs—is stated in Section 3.3 and now priced in Appendix S26.

An objective we had transcribed backwards, and a threshold rule we had not implemented. Adversarial Neuron Pruning trades a natural loss against a loss under adversarial neuron noise. We had put the mixing weight α on the perturbed term rather than the natural one, drew the neuron noise once instead of redrawing it each round, multiplied the whole batch-norm output by the mask where the reference multiplies only the scale—so our "pruned" neuron could not even emit its bias, the mirror image of the BatchNorm defect above—and pruned a fixed global fraction where the method selects its threshold under a clean-accuracy budget. All four are fixed. The correction runs against us in one direction and for us in another: ANP no longer destroys anything, so the twin arm is admissible in all fourteen installations rather than nine, but under its own two-point accuracy budget

it prunes too little to clear this attack ($3/14$, mean residual attack success 0.51) and is excluded on removal success instead. It also removes the one outlier from the surrogate table: the pooled TARE-Z error over all eight defenses is now -0.02 $[-0.17,+0.12]$, against -0.52 $[-5.86, +4.82]$ with the earlier implementation.

An error the twin caught. Our first interpolation-vs-fine-tuning comparison reported a fivefold, sign-reversed gap—an unfair baseline that charged fine-tuning's own cost to removal on one side only (c_{def} for training-side, c_{int} for interpolation). The twin makes it immediate: the same fine-tuning costs *more* on the twin. This is precisely the error the paper attributes to the field, committed once by us and caught by the control we advocate.

Defense grids. The complete grids for both modalities are in Appendix S12. The language model is run at a single fine-pruning operating point (20% of MLP channels); an earlier version reported a $\{10, 20, 30, 40\}\%$ sweep from a single installation, which the defender-draw correction invalidated and which we did not rebuild.

S31 CIFAR-100: THE CROSSOVER AT TWICE THE SWING

The leaderboard gap doubles from CIFAR-10 to CIFAR-100 (Table 3). Two further pre-registrations (Section S43; Table 44 in Section S45) ask whether the causal swing does too, on CIFAR-100 / PreAct-ResNet18 with the benchmark's own CIFAR-100 defense configurations: the same WaNet swap; BPP with the same swap at 5% poisoning (the rate the leaderboard reports for it there); and Blended (Chen et al., 2017) with the step block *added*, as the annealed control (BadNets installs below A0 on CIFAR-100 on the leaderboard).

The swing doubles, on both attacks. Shipped, WaNet's clean twin on CIFAR-100 "gains" -4.54 $[-5.74,-3.34]$ (FT), -2.95 $[-4.41,-1.50]$ (FP) and -4.11 $[-5.17,-3.04]$ pt (NAD) from the defenses; with the cosine schedule FT and NAD show no measurable cost ($+0.34$ $[-0.15,+0.83]$ and $+0.45$ $[-0.10,+1.00]$ pt) and FP costs $+4.41$ $[+3.68,+5.15]$ pt (Table 44). BPP is the same picture, with all six annealed intervals excluding zero: -4.89 $[-6.65,-3.13]$, -3.28 $[-5.54,-1.01]$, -4.62 $[-6.46,-2.79]$ shipped, $+0.54$ $[+0.12,+0.97]$, $+6.20$ $[+5.23,+7.16]$, $+0.62$ $[+0.21,+1.02]$ annealed. The undefended twins move $+6.01$ (WaNet) and $+6.75$ pt (BPP) between schedules, more than twice the CIFAR-10 swing of $+2.68$ and $+2.79$ pt; the shipped twins (65.38, 65.02) sit within 1.4 pt of the benchmark's own never-annealed entries (64.05, 63.96), so this is the leaderboard's CIFAR-100 gap of $+6.3$ pt reproduced causally. Blended, the annealed control, shows the mirror image: its clean-arm cost, pooled over the three defenses, is $+2.76$ $[+2.50,+3.03]$ pt with its shipped cosine schedule and -5.23 $[-7.60,-2.86]$ pt with the step block added (Table 44, "clean arm (pooled)"), and its undefended twin drops from 70.87 to 63.36. The effect follows the three configuration keys, not the attack.

The fixed point holds for the fine-tuning defenses, not for pruning. Over the six CIFAR-100 cells the undefended twins span 8.41 pt; after fine-tuning they span 1.86 pt (69.36–71.23), after NAD 2.07 pt. Fine-pruning does not converge to one level: 5.01 pt across the six cells, with the pruned annealed twins ending *below* the pruned never-annealed ones. Our pre-registered fixed-point test (P11: the widest per-defense range under a third of the undefended range) therefore fails on CIFAR-100 because of FP, and we report it as failed; FT and NAD, the two defenses that re-run a full schedule, satisfy it on their own.

The line also moves the attack. On CIFAR-10 the swap left installation essentially untouched (Table 26). On CIFAR-100 it does not: the shipped, never-annealed WaNet victim installs at ASR 0.767–0.91 across seeds (A0 $1/5$), the annealed one at ≥ 0.97 (A0 $5/5$). The benchmark's published CIFAR-100 WaNet entry at 10% (`results.csv`: accuracy 64.05, ASR 0.977) has the clean accuracy of the shipped victim—our shipped twin sits at 65.38, the annealed one at 71.39—so it was produced under the shipped schedule; its ASR exceeds all five of our shipped seeds, and we could not reproduce that installation rate. Under-training therefore weakens both columns on the harder dataset, and K3 voids the shipped cell's published cost rather than letting us report a cost on an attack that did not install; P6 (the published cost itself) and P7 (interaction invariance) rest on that void cell, so we record both as failed and draw nothing from them. Removal is harder here: A3 clears in a

minority of the fine-pruning cells and in none of the Blended cells (Table 44); no clean-arm number depends on A3.

S32 FT-SAM, SAU AND I-BAU ON THE CROSSOVER CELLS

Table 36: The scheduler crossover on CIFAR-10 / PreAct-ResNet18 for six defenses: clean-arm cost per defense (pt, t-intervals over seeds; negative is a "gain"), the undefended twin, and the A3 row over A0 seeds. The shipped WaNet and BadNets cells use 11 seeds and are the same cells as Table 2; shipped BPP and the three swapped cells use their five pre-registered seeds for all six defenses (SCHED14's eleven-seed FT, FP and NAD swapped cells are in Table 2, so those rows differ slightly). †: SAU on annealed BPP excludes one seed whose defended twin collapsed to 50.6% (post hoc, Section S43); with all five seeds it is $+10.67$ $[-12.25,+33.59]$, $n = 5$.

attack	victim's scheduler	FT	FP	NAD	FT-SAM	SAU	i-BAU	twin base
WaNet	step (shipped)	-1.99 $[-2.30,-1.68]$	-1.68 $[-2.12,-1.24]$	-1.85 $[-2.14,-1.56]$	-2.14 $[-2.46,-1.81]$	-0.15 $[-0.77,+0.47]$	$+0.23$ $[-0.30,+0.75]$	91.50
WaNet	cosine (swapped)	$+0.56$ $[+0.44,+0.68]$	$+0.86$ $[+0.77,+0.95]$	$+1.03$ $[+0.76,+1.30]$	$+1.09$ $[+0.92,+1.26]$	$+2.88$ $[+1.93,+3.82]$	$+5.00$ $[+3.68,+6.31]$	94.23
BadNets	cosine (shipped)	$+0.13$ $[+0.06,+0.21]$	$+1.05$ $[+0.84,+1.25]$	$+0.33$ $[+0.24,+0.42]$	$+0.88$ $[+0.76,+1.00]$	$+2.70$ $[+2.31,+3.08]$	$+5.70$ $[+3.99,+7.42]$	93.71
BadNets	step (swapped)	-2.29 $[-2.59,-1.99]$	-1.80 $[-2.49,-1.12]$	-1.98 $[-2.26,-1.71]$	-2.45 $[-2.84,-2.06]$	-0.58 $[-0.91,-0.25]$	$+0.26$ $[-0.51,+1.03]$	91.16
BPP	step (shipped)	-2.18 $[-2.86,-1.49]$	-1.87 $[-2.48,-1.25]$	-2.00 $[-2.62,-1.39]$	-2.30 $[-3.12,-1.47]$	-0.07 $[-0.80,+0.66]$	$+0.82$ $[-0.13,+1.78]$	91.31
BPP	cosine (swapped)	$+0.55$ $[+0.30,+0.80]$	$+0.97$ $[+0.76,+1.17]$	$+1.11$ $[+0.92,+1.30]$	$+1.10$ $[+0.78,+1.42]$	$+2.41$ $[+1.50,+3.33]^{\dagger}$	$+4.09$ $[+2.70,+5.49]$	94.10
A3 clears (poisoned arm, A0 seeds, six cells)		36/39	23/39	37/39	35/39	37/39	31/39	—
defended twin, range over the four CIFAR-10 cells		0.21	0.70	0.23	0.81	0.73	3.26	3.07

A fourth pre-registration (Section S43) ran three defenses downstream papers compare against—FT-SAM (Zhu et al., 2023a), SAU (Wei et al., 2023), i-BAU—on the same six cells with the benchmark's configurations (14 GPU-hours; the shipped grids' SAU/i-BAU readings predate the pre-registration, which says so).

The flip is not a property of the three old defenses. FT-SAM behaves as FT (-2.14 $[-2.46,-1.81]$ to $+1.09$ $[+0.92,+1.26]$ pt on WaNet, $+0.88$ $[+0.76,+1.00]$ to -2.45 $[-2.84,-2.06]$ pt on BadNets), with defended twins within 0.75 pt of FT's. SAU moves at a different level: within a point of zero on the never-annealed twins (-0.15 $[-0.77,+0.47]$, -0.58 $[-0.91,-0.25]$ pt), a cost of $+2.88$ $[+1.93,+3.82]$ and $+2.70$ $[+2.31,+3.08]$ pt on the annealed ones. i-BAU moves the same way without reversing sign: the never-annealed cells straddle zero ($+0.23$ $[-0.30,+0.75]$, $+0.26$ $[-0.51,+1.03]$, $+0.82$ $[-0.13,+1.78]$ pt) and the annealed ones cost $+5.00$ $[+3.68,+6.31]$, $+5.70$ $[+3.99,+7.42]$ and $+4.09$ $[+2.70,+5.49]$ pt. The BPP cells replicate all three (Table 36; SAU's annealed-BPP cell excludes one collapsed seed, post hoc, with the $n = 5$ value $+10.67$ $[-12.25,+33.59]$ pt in the caption).

Each defense has its own fixed point, and the cost is the distance to it. Across the four WaNet/BadNets cells the undefended twins span 3.07 pt; after FT, NAD and FT-SAM the defended twins span 0.21, 0.23 and 0.81 pt near 93.5%, and after SAU 0.73 pt near 91.4%—a fixed point too, but a lower one. It is not a missing anneal: plain fine-tuning at SAU's own 10^{-4} costs an annealed twin $+0.45$ $[+0.30,+0.60]$ pt over fifty epochs (Section S47), so the $+2.88$ $[+1.93,+3.82]$ pt SAU costs it is the price of its adversarial unlearning, which happens to bring it to about where the never-annealed twins start. That is why SAU is free on those victims and costs two to three points on annealed ones: for SAU as for FT, the published cost is the victim's start minus the defense's level plus what the poison leaves at it. i-BAU has no fixed point (defended range 3.26 pt) and runs against the start: the never-annealed twins finish at 91.27 and 90.90%, the annealed ones at 89.23 and 88.01% (Figure 10); its cost moves with the block by $+4.77$ and $+5.44$ pt, about twice what the block moves the start, so the fixed-point account explains at most half of that swing and the rest we measure and do not explain. The fixed-point statement of Section 3.2 holds for FT, NAD, FT-SAM and SAU, fails for i-BAU and partly for fine-pruning; the sign reversal of Section 4.2 holds with both intervals excluding zero on all twelve CIFAR-10 pairs for FT, FP, NAD and FT-SAM and on the BadNets pair for SAU ($+2.70$ $[+2.31,+3.08]$ to -0.58 $[-0.91,-0.25]$ pt), and on CIFAR-100 (Table 44) for fine-pruning on all three pairs and for FT and NAD on two; for SAU and i-BAU the block changes the size of the cost by two and a half to five and a half points (BPP's SAU swing rests on the † cell), not, in general, its sign. The methods of Section 4.4 beat SAU, FT-SAM and i-BAU as baselines: on a never-annealed victim SAU and i-BAU report near-zero cost, so a method that fine-tunes to its own level reports against them a gain that is the block, not the method.

S33 WHERE THE MILESTONES COME FROM, AND THE BLOCK IN A SECOND TOOLKIT

Seven code bases. The attacks' own releases anneal; BackdoorBench kept their milestones and cut the run to 100 epochs. Table 37 reads the victim's learning-rate schedule off each code base. The WaNet, Input-Aware and BppAttack releases step a `MultiStepLR` at $[100, 200, 300, 400]$ once per epoch over 1000, 600 and 1000 epochs, so their victims first decay at epoch 100 and spend most of training annealed. BackdoorBench carries the same milestones into a 100-epoch run, where the first one would fire only after the last epoch. BackdoorBox's examples (WaNet, BadNets and Blended at $[150, 180]$ of 200 epochs; Input-Aware at the release's milestones over 600) and backdoor-toolbox ($[50, 75]$ of 100 on CIFAR-10) anneal. TrojanZoo's documented attack recipe fine-tunes a pretrained model for 50 epochs at a constant lr 0.01: its scheduler is opt-in (`--lr_scheduler`, off by default), so its poisoning phase ends unannealed too. The coordinate is set three ways across seven code bases, and none of the defense papers we audit reports it (Appendix S1).

Table 37: The victim's learning-rate schedule, read from each code base at the commit given ($\gamma = 0.1$ throughout; the scheduler steps once per epoch in every code base listed). "Decays": whether the rate falls before the last epoch of the run as shipped.

Code base (commit), attack	Victim schedule	Epochs	Decays
WaNet release (`45e8c33`)	`MultiStepLR` $[100, 200, 300, 400]$	1000	yes
Input-Aware release (`1750e4f`), classifier	`MultiStepLR` $[100, 200, 300, 400]$	600	yes
BppAttack release (`f6b6580`)	`MultiStepLR` $[100, 200, 300, 400]$	1000	yes
BackdoorBench (`f02e353`), WaNet, BPP, Input-Aware	`MultiStepLR` $[100, 200, 300, 400]$	100	**no**
BackdoorBench (`f02e353`), LIRA	`MultiStepLR` $[50, 100, 150, 200]$	100	yes
BackdoorBench (`f02e353`), the other twelve	prototype `CosineAnnealingLR`	100	yes
BackdoorBox (`af3afd1c`), WaNet, BadNets, Blended	$[150, 180]$	200	yes
BackdoorBox (`af3afd1c`), Input-Aware	$[100, 200, 300, 400]$	600	yes
backdoor-toolbox (`9d4d909`), CIFAR-10	`MultiStepLR` $[50, 75]$	100	yes
TrojanZoo (`45d680c`), README attack recipe	none (opt-in), lr 0.01 from a pretrained model	50	**no**

The block in BackdoorBox (three registrations). Cutting BackdoorBox's own WaNet run to 100 epochs under its own milestones reproduces the gain with none of BackdoorBench's code. SCHED17 (Section S43) trains BackdoorBox's WaNet, ResNet-18 and training loop (commit `af3afd1c`; only the package `__init__` files trimmed to skip modules with missing dependencies) on CIFAR-10 with its example's settings (lr 0.1, batch 128, 100 epochs), under N, its milestones $[150, 180]$ left in place so no decay fires, or A, milestones $[50, 75]$; five seeds, a poisoned arm at 10% and a twin with ten poisoned images sharing its seed's warping grid. The undefended twin sits at 82.83% under N and 91.44% under A, and the poisoned victim at 79.64% and 90.38% (A0 $5/5$ and $5/5$; A4 $4/5$ and $5/5$, the one N twin above 0.05 at 0.089, below K23's two). Defended by BackdoorBox's own FineTuning (full layers, its default lr 0.001, 10 epochs on a 10% clean split), the twin's tare is -6.10 $[-7.67, -4.52]$ under N and $+0.01$ $[-0.03, +0.04]$ under A. SCHED18, registered while SCHED17's first runs were at epoch 16, adds three BackdoorBox defenses on the same checkpoints: FineTuning at lr 0.01 (-7.12 $[-8.68, -5.57]$ and -0.01 $[-0.10, +0.08]$), NAD (-7.23 $[-8.82, -5.63]$ and $+0.81$ $[+0.67, +0.95]$, the one sign reversal with both intervals excluding zero) and Pruning, the registered negative control with no training phase, which charges both schedules, both intervals excluding zero ($+3.44$ $[+0.83, +6.06]$ and $+1.69$ $[+0.51, +2.87]$; Table 38). **The poisoned arm shows the leaderboard's artifact in a second toolkit:** BackdoorBox's defaults remove its WaNet in no cell (A3 $0/40$ over the four defenses and two schedules), yet on the never-decayed victim FineTuning reports a published cost of -8.24 $[-10.59, -5.90]$, an accuracy "improvement" of the kind Section 4.4's papers report, and under A $+0.19$ $[+0.02, +0.37]$. Predictions P57–P63 (scorecard in Section S43): 5 of seven hold; P59 and P60 fail on their A halves, where fine-tuning at either rate leaves an annealed twin where it was rather than charging it. SCHED19 repeats both registrations on BackdoorBox's own BadNets (3×3 patch, 5%, its example's trigger; Table 38, lower block) and reads the same way: the twin starts 8.94 points lower under N; FineTuning's tare is -5.91 $[-7.92, -3.91]$ under N and $+0.09$ $[-0.05, +0.24]$ under A, NAD's -7.34 $[-9.35, -5.33]$ and $+1.15$ $[+0.51, +1.79]$ (the sign reversal again), Pruning's $+3.46$ $[+1.86, +5.05]$ and $+1.80$ $[+0.68, +2.91]$; A0 $5/5$ and $5/5$, A4 $4/5$ and $5/5$ (the one N twin at 0.051); the defaults remove BadNets in $5/40$ poisoned

cells, Pruning's under N; 4 of its five predictions hold, P65 failing on its A half as P60 did. 21.7 GPU-hours.

Table 38: The scheduler block in BackdoorBox (SCHED17/SCHED18; BadNets block: SCHED19): BackdoorBox's WaNet and BadNets on CIFAR-10 / ResNet-18 under N (milestones $[150, 180]$ in a 100-epoch run: never decays) or A ($[50, 75]$), five seeds; tare and published cost in points with t-intervals (negative is a "gain"); defended twin: clean accuracy after the defense.

	tare (clean arm)		published cost		defended twin (%)	
BackdoorBox defense	N	A	N	A	N	A
WaNet (SCHED17, SCHED18)						
FineTuning, lr 0.001 (BackdoorBox default)	-6.10 [$-7.67, -4.52$]	$+0.01$ [$-0.03, +0.04$]	-8.24 [$-10.59, -5.90$]	$+0.19$ [$+0.02, +0.37$]	88.93	91.43
FineTuning, lr 0.01	-7.12 [$-8.68, -5.57$]	-0.01 [$-0.10, +0.08$]	-9.46 [$-12.08, -6.85$]	-0.06 [$-0.16, +0.04$]	89.95	91.45
NAD	-7.23 [$-8.82, -5.63$]	$+0.81$ [$+0.67, +0.95$]	-9.46 [$-12.17, -6.75$]	$+0.82$ [$+0.42, +1.22$]	90.06	90.63
Pruning (no training phase)	$+3.44$ [$+0.83, +6.06$]	$+1.69$ [$+0.51, +2.87$]	$+3.29$ [$+1.76, +4.81$]	$+2.80$ [$+2.32, +3.28$]	79.39	89.75
Undefended twin (%)	82.83	91.44				
BadNets (SCHED19)						
FineTuning, lr 0.001 (BackdoorBox default)	-5.91 [$-7.92, -3.91$]	$+0.09$ [$-0.05, +0.24$]	-6.57 [$-9.15, -3.98$]	$+0.08$ [$-0.05, +0.20$]	88.68	91.61
FineTuning, lr 0.01	-7.10 [$-9.14, -5.06$]	$+0.05$ [$-0.13, +0.23$]	-7.79 [$-10.23, -5.35$]	$+0.10$ [$+0.03, +0.17$]	89.86	91.65
NAD	-7.34 [$-9.35, -5.33$]	$+1.15$ [$+0.51, +1.79$]	-7.97 [$-10.29, -5.65$]	$+1.09$ [$+0.69, +1.48$]	90.10	90.55
Pruning (no training phase)	$+3.46$ [$+1.86, +5.05$]	$+1.80$ [$+0.68, +2.91$]	$+3.44$ [$+1.88, +5.00$]	$+1.37$ [$+0.67, +2.07$]	79.31	89.91
Undefended twin (%)	82.76	91.70				

S34 ZERO-IMAGE TWINS

Registered as SCHED20 (Appendix S43) before any run started. BackdoorBench refuses a zero poisoning rate (Section 3.1, Appendix S20); for this experiment only, the guard in `utils/backdoor_generate_poison_index.py` is skipped when the environment variable `BB_ALLOW_ZERO_POISON` is set, and a plotting helper that crashed on an epoch with no poisoned sample skips `None` entries (both recorded in `LOCAL_CHANGES.md`); nothing else changes, so a pratio-0 run is the attack's own pipeline with an all-zero poison index (WaNet poisons per batch and draws zero). Three cells, five seeds each, the same seed, recipe and schedule as the paper's ten-image twins: BadNets under the shipped block (cosine), WaNet under the shipped block (never fires) and WaNet with the block removed (cosine); then FT, FP and NAD as everywhere else, and the ten-image twin of the same seed for comparison (Table 39). Registered predictions: P69, the zero-minus-ten undefended accuracy has $|\Delta| \leq 0.3$ points and an interval covering zero in all three cells: ✓; P70, at least 8 of the 9 (cell, defense) tare differences have an interval covering zero and $|\Delta| \leq 0.5$: ✓; P71, all 15 zero-image twins pass A4: ✓. Kill lines: K29 (a one-epoch smoke test, crash only) did not trigger; K30 (the deadline, 2026-09-25 20:00 UTC+8) voided 0 cells; K16 (a twin losing more than 20 points) applied to 0. 5.9 GPU-hours; $60/60$ runs, 0 with a non-zero exit code. Status: final.

Table 39: SCHED20: zero-image twins against the paper's ten-image twins of the same seed (five seeds per cell; $\text{ASR}_{\max}$: the largest attack success any zero-image twin scores, the A4 gate being 0.05). Tare per defense on each twin, and the seed-paired difference with its 95% t-interval.

cell	quantity	zero-image twin	ten-image twin	zero − ten [95% t]	n
BadNets, shipped block (cosine)	clean acc. (%), $\text{ASR}_{\max}$ 0.009	93.82	93.69	$+0.14$ [$-0.12, +0.39$]	5
	tare, FT	$+0.07$ [$-0.05, +0.20$]	$+0.11$ [$+0.00, +0.22$]	-0.04 [$-0.21, +0.13$]	5
	tare, FP	$+1.20$ [$+0.85, +1.54$]	$+1.11$ [$+0.55, +1.67$]	$+0.09$ [$-0.51, +0.68$]	5
	tare, NAD	$+0.38$ [$+0.12, +0.65$]	$+0.26$ [$+0.08, +0.43$]	$+0.13$ [$-0.19, +0.45$]	5
WaNet, shipped block (never fires)	clean acc. (%), $\text{ASR}_{\max}$ 0.015	91.34	91.45	-0.11 [$-0.73, +0.51$]	5
	tare, FT	-2.21 [$-2.84, -1.57$]	-2.01 [$-2.57, -1.44$]	-0.20 [$-0.79, +0.39$]	5
	tare, FP	-1.89 [$-2.55, -1.23$]	-1.74 [$-2.77, -0.72$]	-0.14 [$-1.15, +0.86$]	5
	tare, NAD	-1.93 [$-2.45, -1.41$]	-1.86 [$-2.40, -1.32$]	-0.06 [$-0.54, +0.42$]	5
WaNet, block removed (cosine)	clean acc. (%), $\text{ASR}_{\max}$ 0.009	93.99	94.23	-0.24 [$-0.51, +0.04$]	5
	tare, FT	$+0.51$ [$+0.30, +0.72$]	$+0.56$ [$+0.44, +0.68$]	-0.05 [$-0.21, +0.11$]	5
	tare, FP	$+1.21$ [$+0.89, +1.52$]	$+0.86$ [$+0.77, +0.95$]	$+0.35$ [$+0.03, +0.66$]	5
	tare, NAD	$+0.88$ [$+0.62, +1.15$]	$+1.03$ [$+0.76, +1.30$]	-0.15 [$-0.32, +0.03$]	5

S35 THE CROSSOVER AT 1% AND 5% POISONING

Registered as SCHED21 (Appendix S43) before any run started. The 2×2 of Section 4.2 was run at the benchmark's 10% rate; here the victims are re-trained at pratio 0.05 and 0.01, five seeds

each, under the shipped block and under the swapped one (WaNet: the shipped `MultiStepLR` that never fires against `CosineAnnealingLR`; BadNets: the shipped `CosineAnnealingLR` against a `MultiStepLR` whose first milestone is epoch 100), then FT, FP and NAD as everywhere else; a twin has no rate, so the tare is the paper's ten-image twin of the same seed and schedule, not re-run (Table 40). Registered predictions: P72, the annealed victim starts at least one point higher in at least 4 of 5 seeds for each (attack, rate): $\times$; P73, on WaNet the published cost of FT, FP and NAD is negative under the shipped block and higher under the swapped one, the difference's interval excluding zero, in at least 5 of 6 (defense, rate) pairs: $\times$; P74, the same on BadNets: $\times$; P75, WaNet at 1% installs (A0 in at least 3 of 5 seeds) under both schedules: $\times$. Kill lines: K31 (A0 failing in 3 or more of 5 seeds voids that victim cell): WaNet at 1% under the shipped block, void (K31: A0 fails in $3/5$); under the swapped block, counted; BadNets at 1% under the shipped block, void (K31: A0 fails in $5/5$); under the swapped block, void (K31: A0 fails in $5/5$); every 5% cell counted (A0 $4/5$, $5/5$, $4/5$, $4/5$). K32 (the deadline, 2026-09-25 20:00 UTC+8) voided 0 cells. A prediction whose test arm a kill line voids counts as failed, so the four predictions, each spanning both rates, are recorded as failed although their 5% halves hold (P72 $2/4$, P73 $3/6$, P74 $3/6$). A0 over all victims $23/40$, A3 over all defended victims $29/69$; $40/40$ victims, 15.1 GPU-hours. Status: final.

Table 40: SCHED21: the scheduler crossover at 5% and 1% poisoning. Per (attack, rate, schedule): A0 (seeds installing), the undefended victim's accuracy, the published cost of FT, FP and NAD (mean over A0 seeds, 95% t-interval) and the tare on the same-seed ten-image twin; the third line of each block is the annealed-minus-not-annealed difference.

				undefended	published cost [95% t]			tare (same-seed twin)		
attack	rate	schedule	A0	acc. (%)	FT	FP	NAD	FT	FP	NAD
WaNet	5%	shipped (`MultiStepLR`, never fires)	4/5	89.50	-4.11 [-7.05, -1.17]	-4.12 [-7.13, -1.11]	-3.97 [-6.91, -1.03]	-2.01 [-2.57, -1.44]	-1.74 [-2.77, -0.72]	-1.86 [-2.40, -1.32]
		swapped (`CosineAnnealingLR`)	5/5	91.99	-1.45 [-3.86, $+0.96$]	-1.37 [-3.62, $+0.88$]	-1.00 [-3.27, $+1.27$]	$+0.56$ [$+0.44$, $+0.68$]	$+0.86$ [$+0.77$, $+0.95$]	$+1.03$ [$+0.76$, $+1.30$]
		annealed − not annealed		$+2.49$ [$+0.65$, $+4.33$]	$+2.77$ [$+0.34$, $+5.20$]	$+2.89$ [$+0.36$, $+5.41$]	$+3.07$ [$+0.67$, $+5.48$]			
WaNet	1%	shipped (`MultiStepLR`, never fires)	2/5	90.32	–	–	–	-2.01 [-2.57, -1.44]	-1.74 [-2.77, -0.72]	-1.86 [-2.40, -1.32]
		swapped (`CosineAnnealingLR`)	4/5	92.42	-1.22 [-3.16, $+0.73$]	-0.98 [-2.86, $+0.91$]	-0.77 [-2.79, $+1.24$]	$+0.56$ [$+0.44$, $+0.68$]	$+0.86$ [$+0.77$, $+0.95$]	$+1.03$ [$+0.76$, $+1.30$]
		annealed − not annealed		$+2.10$ [$+0.94$, $+3.26$]	$+2.71$ [$+1.31$, $+4.11$]	$+2.80$ [$+2.55$, $+3.05$]	$+3.07$ [$+0.33$, $+5.80$]			
BadNets	5%	swapped (`MultiStepLR`, first milestone 100)	4/5	90.25	-2.70 [-3.92, -1.49]	-2.82 [-4.01, -1.62]	-2.46 [-3.78, -1.13]	-2.29 [-2.59, -1.99]	-1.80 [-2.49, -1.12]	-1.98 [-2.26, -1.71]
		shipped (`CosineAnnealingLR`)	4/5	92.51	$+1.36$ [$+1.11$, $+1.62$]	$+0.07$ [-0.30, $+0.44$]	$+2.26$ [$+1.76$, $+2.76$]	$+0.11$ [$+0.00$, $+0.22$]	$+1.11$ [$+0.55$, $+1.67$]	$+0.26$ [$+0.08$, $+0.43$]
		annealed − not annealed		$+2.26$ [$+1.14$, $+3.38$]	$+3.86$ [$+1.71$, $+6.01$]	$+2.77$ [$+0.39$, $+5.15$]	$+4.68$ [$+1.84$, $+7.51$]			
BadNets	1%	swapped (`MultiStepLR`, first milestone 100)	0/5	91.01	–	–	–	-2.29 [-2.59, -1.99]	-1.80 [-2.49, -1.12]	-1.98 [-2.26, -1.71]
		shipped (`CosineAnnealingLR`)	0/5	93.45	–	–	–	$+0.11$ [$+0.00$, $+0.22$]	$+1.11$ [$+0.55$, $+1.67$]	$+0.26$ [$+0.08$, $+0.43$]
		annealed − not annealed		$+2.45$ [$+1.77$, $+3.12$]	–	–	–			

S36 ANP UNDER A DEPLOYER-SIDE THRESHOLD RULE

Registered as SCHED22 (Appendix S43), an analysis-only registration written before the analysis ran. BackdoorBench's ANP scans pruning thresholds and keeps the one with the smallest *test-set* attack success within a clean-accuracy budget ($|Q - Q_{\text{ori}}|/Q_{\text{ori}} < 0.1$; `defense/anp.py`), an oracle rule the paper discloses at every ANP cell (†). Every ANP run in the archive saved its full sweep (`threshold_df.csv`), so each cell can be re-read under a rule that never sees attack success: *rule D*, the largest threshold (the most pruning) whose clean accuracy stays within the same budget. Rule O, the framework's, is reproduced from the sweeps in $225/225$ runs to 10^{-6} (the sweep's row 0 is not the unpruned model—neurons whose mask prints as 0.0000 are already pruned there—so rule O takes the unpruned accuracy from `attack_df_summary`, as the framework does; rule D keeps row 0 as registered, and a variant on the framework's budget changes no verdict). Registered predictions: P80, BadNets tare under rule D positive with interval excluding zero: ✓ ($+6.44$ [$+5.11$, $+7.77$]); P81, the two rules' BadNets tares within one point: × (D − O $+3.95$ [$+2.35$, $+5.55$]); P82, WaNet tare of the same sign under both rules: ✓ (O $+1.78$ [$+0.52$, $+3.04$], D $+6.79$ [$+5.88$, $+7.71$]). The mechanism is visible in the thresholds: on the poisoned arm the two rules pick the same threshold in $97/110$ runs, on the twin arm in $35/110$—on a model with no backdoor the oracle rule picks the minimum of attack-success noise and often barely prunes, so it makes ANP look cheap on the twin; the paired contrast on WaNet moves from $+4.95$ [$+1.96$, $+7.93$] to -0.11 [-2.05, $+1.82$]. Of the seven other attacks, five have a higher tare under rule D and two (Trojan, SIG) the same threshold under both rules (Table 41); the threshold counts run over the poisoned and twin arms of the registered and descriptive cells and of the dose-response cells that share a twin; runs without a sweep at collection time (0) are not counted.

Table 41: SCHED22: ANP's tare under the framework's threshold rule (O: smallest test-set attack success within the budget) and under a deployer-side rule (D: most pruning within the budget, attack success unseen), seed-paired, 95% t-intervals; A3 under each rule; n seeds. The two registered cells first, the other Table 4 attacks descriptive.

cell	tare, rule O [95% t]	tare, rule D [95% t]	D − O [95% t]	A3, O	A3, D	n
BadNets (registered)	$+2.49$ [$+0.39$, $+4.60$]	$+6.44$ [$+5.11$, $+7.77$]	$+3.95$ [$+2.35$, $+5.55$]	$11/11$	$11/11$	11
WaNet (registered)	$+1.78$ [$+0.52$, $+3.04$]	$+6.79$ [$+5.88$, $+7.71$]	$+5.01$ [$+3.41$, $+6.61$]	$8/8$	$8/8$	11
Blended	$+3.86$ [$+1.01$, $+6.72$]	$+5.95$ [$+4.01$, $+7.90$]	$+2.09$ [-0.02, $+4.20$]	$6/8$	$6/8$	8
LF	$+3.25$ [$+0.17$, $+6.33$]	$+5.20$ [$+3.76$, $+6.64$]	$+1.95$ [-0.37, $+4.27$]	$5/5$	$5/5$	5
FTrojanNN	$+1.43$ [$+0.42$, $+2.43$]	$+5.78$ [$+3.31$, $+8.25$]	$+4.35$ [$+1.71$, $+6.99$]	$5/5$	$5/5$	5
Refool	$+4.72$ [$+1.11$, $+8.34$]	$+6.18$ [$+3.54$, $+8.82$]	$+1.45$ [-1.02, $+3.93$]	$5/5$	$5/5$	5
Trojan	$+6.53$ [$+5.53$, $+7.52$]	$+6.53$ [$+5.53$, $+7.52$]	0.00 (same row, every seed)	$5/5$	$5/5$	5
SIG	$+7.27$ [$+5.52$, $+9.02$]	$+7.27$ [$+5.52$, $+9.02$]	0.00 (same row, every seed)	$5/5$	$5/5$	5
BPP	$+4.28$ [$+1.80$, $+6.75$]	$+6.55$ [$+4.61$, $+8.49$]	$+2.27$ [$+0.28$, $+4.26$]	$5/5$	$5/5$	5

S37 THE EIGHT-DEFENSE CROSSOVER

Registered as SCHED23 (Appendix S43) before any run started; no new victim or twin was trained. The defenses missing from the swapped and shipped arms of the crossover (NC, ANP, ABL on WaNet swapped and BadNets swapped; six on BPP swapped) were run on the existing arms so that WaNet (11 seeds), BadNets (swapped: 5) and BPP (5) carry all eight BackdoorBench defenses under both schedules (Table 42). Registered predictions: P76, ANP's tare on WaNet positive under both schedules: ✓; P77, ABL destroys most twins under both schedules (WaNet never $7/11$; WaNet cosine $7/11$; BPP never $3/5$; BPP cosine $5/5$): ✓; P78, NC's detector fires on at most half of the cosine WaNet twins ($6/11$) and its repair costs them a positive amount where it fires ($+3.31$): ×; P79, the benchmark's own DER ordering of the eight defenses on WaNet changes between the shipped and the patched victim in at least one (fine-tuning-family, other) pair: ✓ (FP/I-BAU, FP/SAU; Table 43). Kill lines: K33 voided 0 (cell, defense) pairs; K34 (the deadline, 2026-09-26 09:00 UTC+8) voided 0; K16 applied to 2 non-ABL twins (WaNet cosine SAU: s5; BPP cosine SAU: s2), excluded as failure records. $146/146$ registered runs plus the make-up runs of FT-SAM, I-BAU and SAU on six WaNet-cosine seeds the registration wrongly assumed present (its outcome record), 27.1 GPU-hours. Status: harvest before the deadline, every registered arm complete.

Table 42: SCHED23: the tare of all eight defenses on each arm of the crossover (mean over seeds, 95% t-interval; WaNet 11 seeds, BadNets swapped and BPP 5; † policy-level A2; ♮ ABL is a failure record, ♭k/n twins destroyed; NC printed as twins fired | cost where fired; –: not harvested by the deadline).

attack	schedule	FT	FP	NAD	NC†	ANP†	ABL†♮	I-BAU	SAU
WaNet	never (shipped)	-1.99 ± 0.31	-1.68 ± 0.44	-1.85 ± 0.29	$3/11 \mid -1.30$	$+1.78$ ± 1.26	$+27.77$ ± 6.49♭7/11 destroyed	$+0.23$ ± 0.53	-0.15 ± 0.62
WaNet	cosine (swapped)	$+0.51$ ± 0.08	$+0.86$ ± 0.12	$+0.99$ ± 0.13	$6/11 \mid +3.31$	$+3.96$ ± 1.74	$+30.45$ ± 6.47♭7/11 destroyed	$+4.76$ ± 0.69	$+2.61$ ± 0.44♭1/11
BadNets	cosine (shipped)	$+0.13$ ± 0.07	$+1.05$ ± 0.20	$+0.33$ ± 0.09	$5/11 \mid +2.71$	$+2.49$ ± 2.10	$+33.20$ ± 4.06♭11/11 destroyed	$+5.70$ ± 1.71	$+2.70$ ± 0.39
BadNets	step (swapped)	-2.29 ± 0.30	-1.80 ± 0.69	-1.98 ± 0.28	$4/5 \mid -1.24$	$+1.57$ ± 1.76	$+30.72$ ± 8.70♭4/5 destroyed	$+0.26$ ± 0.77	-0.58 ± 0.33
BPP	never (shipped)	-2.07 ± 0.75	-1.82 ± 1.07	-1.84 ± 0.64	$0/5 \mid$ —	$+4.28$ ± 2.48	$+29.83$ ± 12.18♭3/5 destroyed	$+0.71$ ± 1.67	-0.60 ± 0.96
BPP	cosine (swapped)	$+0.58$ ± 0.33	$+0.94$ ± 0.19	$+0.96$ ± 0.31	$0/5 \mid$ —	$+4.17$ ± 2.19	$+32.46$ ± 12.42♭5/5 destroyed	$+5.28$ ± 1.53	$+2.49$ ± 0.80♭1/5

Table 43: SCHED23: the benchmark's DER (victim-referenced, clipped; mean over the A0 seeds harvested, which needs only the poisoned arm) for the eight defenses on the shipped and the patched victims, with the rank in parentheses; –: not harvested by the deadline.

attack	schedule	FT	FP	NAD	NC†	ANP†	ABL†♮	I-BAU	SAU
WaNet	never (shipped)	97.5 (2)	95.8 (5)	97.7 (1)	80.0 (7)	95.6 (6)	67.8 (8)	96.5 (4)	97.2 (3)
WaNet	cosine (swapped)	99.1 (1)	98.6 (3)	99.0 (2)	85.4 (7)	96.7 (6)	68.3 (8)	97.0 (5)	98.1 (4)
BadNets	cosine (shipped)	96.3 (3)	97.1 (1)	95.8 (4)	83.6 (8)	94.3 (6)	92.9 (7)	94.8 (5)	96.4 (2)
BadNets	step (swapped)	85.9 (8)	93.0 (5)	88.8 (6)	88.4 (7)	95.4 (3)	94.6 (4)	97.1 (2)	97.8 (1)
BPP	never (shipped)	98.4 (2)	81.3 (7)	98.5 (1)	89.0 (6)	96.3 (4)	56.5 (8)	94.6 (5)	98.1 (3)
BPP	cosine (swapped)	98.2 (1)	80.5 (7)	98.1 (2)	97.8 (3)	96.6 (5)	55.2 (8)	94.4 (6)	97.0 (4)

S38 Sources and hashes for the leaderboard analysis

The under-convergence gap per (dataset, backbone) is Table 3 in the main text.

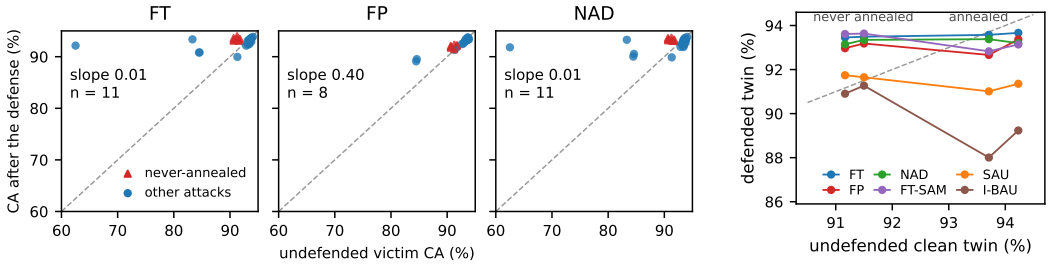
The rendered CIFAR-10 leaderboard (S1), the site's previous and current `results.csv` (S2, S3), Tables 3/4/13 of the 2022 paper (S4) and Table III of the 2025 paper (S5) were downloaded on 2026-09-12; SHA-256 sums are in the supplement. All runs use BackdoorBench at commit `f02e353` (2025-07-04) with seven local changes, listed in `supplement/sched/LOCAL_CHANGES.md`: six NumPy-2 compatibility renames (`np.round_`, `np.infty`; `np.float` in the patch transform and in the Trojan attack, made for SCHED10; `np.float` in the FTrojanNN transform and `np.lib.pad` in the Refool attack, made for SCHED15) and a label-transform option no run behind this paper selects; none alters a number in this paper. We do not know the commit or seed behind the public CIFAR-100 WaNet entry, whose ASR (0.977) we could not reproduce under the shipped configuration (Section S31). S1 lists ten attacks on CIFAR-10, eight at 10% (LC and SIG have no entry there) and carries BPP; S3 lacks BPP at 10% and carries LIRA, Blind, LC and SIG. The sign counts, permutation tests and the $13/15$ vs. $0/25$ DER count of Section 4.4 use S1; Figure 2 and Section S39. S2 and S3 agree with S1 in 226 of 252 CIFAR-10 / PreAct-ResNet18 cells; the 26 mismatches are LC cells re-run for the current site and one LF/ABL cell at 0.5%, none in a cell used in Section 4.4. Table 3 excludes the 10% cells because LC and SIG are annealed but heavily damaged at that rate, which was stated before the four-dataset table was computed.

S39 Leaderboard fixed point, all rates

For each defense and poisoning rate on CIFAR-10 and CIFAR-100 / PreAct-ResNet18, the number of attacks with a public entry, the least-squares slope of after-defense accuracy on before-defense accuracy across those attacks, and the two ranges. Entries recorded as -100 (failed runs) are dropped. The last three columns restrict the 10% CIFAR-10 cells to the eight attacks all three defenses cover, to the ten attacks without the LIRA victim (which alone spans most of the undefended range), and to the healthy annealed attacks (neither never-annealed nor damaged by their own poisoning: 5 attacks spanning 2.1 points). On that last subset the slope is $+1.26$ (SE 0.10) for FT and $+1.08$ (SE 0.20) for NAD, and the defended spread (2.82 points under FT) is not smaller than the undefended one (2.15): on healthy annealed victims the defense leaves the start near where it was (the healthy/damaged split was made after reading the leaderboard). The slope on this subset leans on BadNets, its lowest start: without it the other four span 0.59 points and the slopes are $+0.86$ (FT) and $+0.14$ (NAD), too

narrow a spread to pin a slope. The fixed point is a property of the victims that start low, and the main text says so wherever it is stated; the twin-arm evidence of Section 3.2 and Figure 2 (four cells 3.07 points apart returned to within 0.21 by FT) is its direct measurement. Over every (dataset, backbone, poisoning-rate) cell with at least four attacks outside GTSRB, the compression criterion ($|\text{slope}| < 0.25$ and a defended range under a third of the undefended one) holds in $18/41$ FT, $11/39$ NAD and $0/9$ fine-pruning cells; it can also fail where the victims start close together, since a third of a small range is within seed-to-seed noise.

dataset	defense	rate	n	slope	undefended range	defended range	common n	common-set slope	slope w/o LIRA	healthy-annealed slope (n)
CIFAR-10	FT	10%	11	+0.01	31.0	3.5	8	+0.18	+0.06	+1.26 (5)
CIFAR-10	FT	5%	5	−0.20	2.3	0.9	—	—	—	—
CIFAR-10	FT	1%	9	+0.01	3.2	0.5	—	—	—	—
CIFAR-10	FT	0.5%	5	−0.01	3.2	0.8	—	—	—	—
CIFAR-10	FP	10%	8	+0.40	9.0	4.1	8	+0.40	+0.40	+0.85 (4)
CIFAR-10	FP	5%	8	+0.70	2.2	1.6	—	—	—	—
CIFAR-10	FP	1%	8	+0.50	3.2	1.5	—	—	—	—
CIFAR-10	FP	0.5%	4	+0.50	3.2	1.8	—	—	—	—
CIFAR-10	NAD	10%	11	+0.01	31.0	3.4	8	+0.22	+0.07	+1.08 (5)
CIFAR-10	NAD	5%	5	−0.21	2.3	0.6	—	—	—	—
CIFAR-10	NAD	1%	9	−0.07	3.2	1.1	—	—	—	—
CIFAR-10	NAD	0.5%	5	−0.06	3.2	0.7	—	—	—	—
CIFAR-100	FT	10%	7	+0.01	65.3	3.2	—	—	—	—
CIFAR-100	FT	5%	7	−0.16	6.1	1.8	—	—	—	—
CIFAR-100	FT	1%	6	+0.06	7.4	0.8	—	—	—	—
CIFAR-100	FT	0.5%	6	+0.15	6.4	1.2	—	—	—	—
CIFAR-100	FT	0.1%	6	+0.08	6.1	0.7	—	—	—	—
CIFAR-100	FP	10%	4	+0.45	5.2	2.4	—	—	—	—
CIFAR-100	FP	5%	6	+0.32	6.1	2.1	—	—	—	—
CIFAR-100	FP	1%	5	+0.48	7.1	3.7	—	—	—	—
CIFAR-100	FP	0.5%	5	+0.56	6.4	3.6	—	—	—	—
CIFAR-100	FP	0.1%	5	+0.62	5.8	3.7	—	—	—	—
CIFAR-100	NAD	10%	7	+0.01	65.3	3.8	—	—	—	—
CIFAR-100	NAD	5%	7	−0.10	6.1	1.5	—	—	—	—
CIFAR-100	NAD	1%	6	+0.12	7.4	1.6	—	—	—	—
CIFAR-100	NAD	0.5%	6	+0.17	6.4	1.3	—	—	—	—
CIFAR-100	NAD	0.1%	6	+0.15	6.1	1.2	—	—	—	—



(a) (b)

Figure 10: The level per defense. **(a)** Every attack at every rate on the public CIFAR-10 / PreAct-ResNet18 leaderboard (the site's `results.csv`): accuracy after FT, NAD and FP against before; slopes at 10%: 0.01 and 0.01 over all eleven attacks, 0.18/0.22 on the eight all three cover, against +1.26 and +1.08 on the healthy annealed five (grouped after reading the leaderboard). Triangles: the never-annealed three. **(b)** The same on the never-poisoned twins we trained, one line per defense over the four WaNet/BadNets cells: FT, NAD and FT-SAM return every twin to 92.8–93.7%, SAU to near 91.4%; I-BAU has no level of its own.

Grouping sensitivity (post hoc). The healthy annealed five of Figure 2 were grouped after reading the leaderboard; any cutoff on undefended accuracy between 84.6 and 91.3% selects the same five. Over every annealed attack at 10% on the `results.csv` roster, the damaged four (SIG, LC, LIRA, Blind) included, the FT and NAD slopes are −0.002 and −0.001: those four start low and are returned to the level as the never-annealed victims are, so the healthy-five contrast is a statement about victims that already start at the level, not about annealing as such; the clip then fires in $16/40$ annealed (attack, defense) cells on that roster, none of them a healthy victim's (the rendered site's own 10% page carries no damaged attack, hence its $0/25$).

S40 THE CORRECTED REFERENCE

Removing the clip from its accuracy term is not enough: $-\Delta Q$ without the $\max(0, \cdot)$ would reward a negative cost rather than expose it—on the shipped WaNet grid (A0-gated published cost -3.07 $[-3.42,-2.73]$ pt; -3.05 $[-3.37,-2.73]$ A3-conditioned in Table 26) the unclipped term adds $+1.54$ DER points, where the clean-arm cost of $+0.78$ $[+0.70,+0.87]$ pt would subtract -0.39. The clip hides the artifact; unclipping alone amplifies it in the headline number—a nonlinearity manufacturing a qualitative claim in either direction (Schaeffer et al., 2023). DER's accuracy term should instead be referenced to a clean model trained from the prototype recipe—one per (dataset, backbone, seed), sharing no scheduler key with any attack—as $-[Q(\theta_{\text{clean}}) - Q(D\theta_{\text{pois}})]$. This reference does not depend on the victim's schedule, so a never-annealed victim can no longer be booked as a gain, and the defense's own cost on a clean model stays in the score: i-BAU's $+5.00$ $[+3.68,+6.31]$ and SAU's $+2.88$ $[+1.93,+3.82]$ pt on annealed twins (Table 36) are paid by every model those defenses touch, and a defended-clean reference $-[Q(D\theta_{\text{clean}}) - Q(D\theta_{\text{pois}})]$ would cancel them (that bracket is the published cost minus the clean-arm cost plus the start gap). With the prototype reference (here the same seed's BadNets-cosine twin) the bracket is $+0.59$ $[+0.44,+0.74]$ shipped and $+0.77$ $[+0.66,+0.88]$ pt annealed on WaNet, $+0.57$ $[+0.29,+0.85]$ and $+0.89$ $[+0.55,+1.23]$ pt on BPP—within half a point across schedules, a non-rejection at five seeds—and $+3.03$ $[+2.89,+3.16]$ pt on BadNets with cosine, $+1.28$ $[+1.13,+1.43]$ pt‡ with the block. The prototype clean model serves every attack, including Input-Aware; where the fixed point fails (i-BAU, fine-pruning on CIFAR-100) the per-attack twin remains the right flag. $Q(D\theta_{\text{clean}})$ and the signed, unclipped ΔQ belong beside the score as flag columns (the clean-arm cost, and the twin-referenced bracket of Table 26), so that a reader can tell the three parts of the cost apart while the score that gets copied keeps all three.

S41 LEADERBOARD RE-RANKING UNDER THE CORRECTED REFERENCE

For each defense with public PreAct-ResNet18 entries, mean DER over the attacks shared by all defenses in the panel (CIFAR-10: every defense with at least five entries at 10%, then the fine-tuning family at 10% and 5%; CIFAR-100 (C100): every defense with at least four entries, prototype level 70.75%, the mean 0.1%-poisoning accuracy of the healthy annealed attacks), under the benchmark's DER and under a DER whose accuracy term is referenced to the benchmark's clean PreAct-ResNet18 model (93.65%, Wu et al., 2025; the 93.9% of the 2022 paper gives the same order). Original order at 10%: ANP > ABL > i-bau > FT > NC > FP > NAD > CLP > spectral > d-br > DBD > AC; prototype-referenced: ANP > ABL > i-bau > FT > NAD > FP > NC > CLP > spectral > d-br > AC > DBD.

defense	rate	attacks	DER (original)	rank	DER (prototype-referenced)	rank
FT	10%	4	74.25	4	73.75	4
FP	10%	4	71.89	6	70.93	6
NAD	10%	4	71.85	7	71.22	5
NC	10%	4	72.20	5	70.70	7
ANP	10%	4	94.69	1	93.12	1
i-BAU	10%	4	88.62	3	87.20	3
ABL	10%	4	90.26	2	88.69	2
CLP	10%	4	53.44	8	51.88	8
DBD	10%	4	49.50	11	47.94	12
AC	10%	4	49.45	12	47.97	11
Spectral	10%	4	50.70	9	49.22	9
D-BR	10%	4	50.28	10	48.72	10
FT (C100)	10%	4	73.64	6	72.55	6
FP (C100)	10%	4	60.74	7	59.16	7
NAD (C100)	10%	4	73.99	5	72.85	5
NC (C100)	10%	4	96.23	1	94.97	1
ANP (C100)	10%	4	84.12	4	82.45	4
i-BAU (C100)	10%	4	86.24	3	84.73	3
ABL (C100)	10%	4	90.16	2	88.49	2
CLP (C100)	10%	4	46.50	11	44.83	11
DBD (C100)	10%	4	44.43	12	42.76	12
AC (C100)	10%	4	48.52	10	46.85	10
Spectral (C100)	10%	4	50.99	8	49.62	8
D-BR (C100)	10%	4	42.13	13	40.46	13
D-ST (C100)	10%	4	49.19	9	47.52	9
FT (C100)	5%	4	64.41	7	63.83	7
FP (C100)	5%	4	63.98	8	62.90	8
NAD (C100)	5%	4	65.46	6	64.81	6
NC (C100)	5%	4	91.83	1	91.12	1
ANP (C100)	5%	4	74.95	4	73.65	4
i-BAU (C100)	5%	4	74.18	5	73.19	5
ABL (C100)	5%	4	85.81	2	84.52	2
CLP (C100)	5%	4	60.46	9	59.17	9
DBD (C100)	5%	4	77.61	3	76.31	3
AC (C100)	5%	4	51.47	11	50.18	11
Spectral (C100)	5%	4	50.52	12	49.67	12
D-BR (C100)	5%	4	42.93	13	41.63	13
D-ST (C100)	5%	4	55.45	10	54.15	10
FT	10%	6	75.88	4	75.46	4
FP	10%	6	72.78	5	71.94	5
NAD	10%	6	76.83	3	76.33	3
NC	10%	6	80.95	2	79.83	2
i-BAU	10%	6	85.34	1	84.11	1
FT	5%	3	59.88	3	59.72	3
FP	5%	3	66.11	2	65.51	2
NAD	5%	3	56.42	4	56.22	4
NC	5%	3	50.00	5	49.28	5
i-BAU	5%	3	83.24	1	82.55	1

S42 LEADERBOARD COST TABLE

Per-(attack, defense) ΔQ at 10% and 5% poisoning, with sign counts by group and exact permutation tests, are printed by `leaderboard/leaderboard_artifact_test.py`; the markdown report it writes is included verbatim in the supplement.

S43 PRE-REGISTRATIONS

All twenty-five reported pre-registrations are included verbatim (for double-blind review, zone labels, the GPU model name and one host path in the scheduler study's files were normalized after registration, to "UTC+8", "GPU" and "<server>"; no other text changed)—the scheduler study's (`supplement/sched/prereg`; SCHED3b as an addendum inside the SCHED3 file), the eleven-seed WaNet grid's (`supplement/prereg`) and the ranking protocol (`supplement/configs`)—with timestamps, predictions with priors and kill lines; the timestamps are those typed in each file, and for SCHED11 onward also our file system's record of when the file was written, which the supplement's zip (modification times only) does not carry; none is an external registry, and one arm of the CIFAR-10 crossover was already known when it was written (stated below). Files: the CIFAR-10 crossover (WaNet $\times$ BadNets); the BPP / Input-Aware extension; the CIFAR-100 crossover (WaNet, with Blended as the annealed control); and its BPP addendum (SCHED3b, at the poisoning rate the leaderboard reports for BPP on CIFAR-100); and SCHED4, the three further defenses on the existing records (Section S32), whose one deviation is a single SAU run (BPP cosine, seed 4) whose defended twin collapsed to 50.6% and is excluded as a failed run—a post-hoc judgement, not a pre-registered rule, so the † note of the Table 36 caption gives the interval with it included ($+10.67$ $[-12.25,+33.59]$, $n = 5$) next to the $n = 4$ interval printed in the cell. Every deviation is logged in the file it belongs to and stated in the text (one exception: the SCHED9 and SCHED10 header-time corrections, stated only below): the BPP twin reset from ten images to one on CIFAR-10 (prescribed by SCHED2's K3, Section S46); rule K3, registered before any CIFAR-100 run as voiding a whole cell whose attack installs in a minority of seeds, and its disposition, written after seed 0 of the shipped WaNet cell returned a low installation rate and before seeds 1–4 existed: void the poisoned arm and report the clean arm as a declared deviation; the failed predictions P6 and P7 (published cost and interaction invariance of the K3-voided CIFAR-100 cell) and P11 (four-cell fixed point on CIFAR-100), reported as failed in Section S31; and the one-image twin for Blended, which follows the convention established for BPP on CIFAR-10. Six decisions were taken after data existed. Two are rules written before their outcome was determined: the BPP twin reset (before the one-image twins were run) and K3's disposition (after seed 0 of the shipped CIFAR-100 WaNet cell returned an ASR below A0, before seeds 1–4, whose installation decided whether it fired; timestamps in the SCHED3 file). The third was taken after seeing its outcome: the exclusion of the SAU run on the BPP-cosine twin of seed 4, implemented in `build_sched4.py` as "a defended twin that loses more than 20 pt is a failed run"; the threshold was written after the run had been seen, it affects only that twin (the seed's poisoned arm is kept), and it is the only reason the SAU cell for annealed BPP reads $+2.41$ $[+1.50,+3.33]$ ($n = 4$) rather than $+10.67$ $[-12.25,+33.59]$ ($n = 5$). Every sentence that leans on that cell is marked †. The fourth and fifth were also taken after their data: TSBD's operating point, fixed at the paper's printed default after the first harvest had taken the sweep's last value (Section S53), and SIG's twin, left unreset after it failed A4 although SCHED9's K14 allowed one one-image reset (Section S49). No other number in the tables was chosen after seeing data; the sixth decision was a grouping: Figure 2's healthy/damaged split of the leaderboard's annealed attacks (Section S39). The eleven-seed shipped grids and their SAU / i-BAU readings predate the pre-registrations that use them, which state so: the CIFAR-10 crossover and SCHED4 were pre-registered with one arm already known, SCHED2, SCHED3 and SCHED3b with both arms unrun, and SCHED13 reuses three seed-0 BadNets victims (at 1, 5 and 20%, with their FT and NAD runs) that predate it, as its file states; the 5% one is the seed that fails A0 there. The eleven-seed shipped grids predate the pre-registrations (they were run for an earlier study); the admissibility thresholds (A0 0.90, A3/A4 0.05) and the ten-image twin convention were fixed with them. SCHED5 (the FT budget sweep, Section S47), SCHED6 (the VGG19-BN crossover, Section S48) and SCHED7 (TSBD and CGD measured in the WaNet 2×2, Section S53) were added in that order, each written before its runs. SCHED9 (SIG) and SCHED10 onward followed, each written before its new runs. For SCHED11–SCHED15 the evidence is the file system's write time, which precedes each launcher's first run (their typed header times are 4 to 17 minutes later and are not the evidence); SCHED16's file was written at 01:34 UTC on 2026-09-23, the minute its header states, before its launcher; SCHED17's at 03:02 UTC on 2026-09-24, six minutes before its launcher started, and SCHED18's and SCHED19's at 03:14 and 03:35 UTC, before SCHED17's first readout at 03:39 (SCHED18's header time was corrected by two minutes at 03:15, recorded in its file); for SCHED9 and SCHED10 it is only our command log, not in the supplement, which shows each file written seconds and three minutes before its first run, while their typed header times, corrected after launch, read 1 and 12 minutes earlier than the files were written. One exception to

"before its new runs": seven of SCHED11's seed-0 BadNets cells predate it—CLP and NPD on the poisoned arm are runs of 2026-08-28 under the same configuration, and the five defenses on the twin come from the timing smoke test whose RNP and D-BR readings P39's prior quotes. One deviation: D-ST's registered override (adding only `client_optimizer`) failed on its first arm, which under SCHED11 drops D-ST; we then widened the override to the fine-tuning defaults and ran it twice on the seed-0 BadNets twin (a trial stopped at epoch 13 and a launcher run stopped by hand at epoch 115 of its 200, about 2.1 GPU-hours in all) before dropping it for cost; no D-ST number is printed. SCHED15 was registered as adding two leaderboard attacks; FTrojanNN and Refool are BackdoorBench attacks the public leaderboard does not list, and the paper says so. A prediction whose test arm a kill line voids is counted as failed (SCHED2's P3, P4 and the Input-Aware half of P5; SCHED3's P1, P3 and P4 under K3 as registered, which held as its declared deviation; P6; P7; P45; P46's ratio clause; P47). The scorecard below covers the scheduler study. The ranking protocol's three criteria are scored in Section S19 (criterion (i) fails), and the eleven-seed WaNet grid's three predictions all held: P1, its A3-passing published cost negative with the interval excluding zero (-1.00 $[-1.53, -0.47]$, Table 29's convention); P2, no reordering (2 of 20 testable pairs reverse, 0 survives, Section S19); P3, ABL failing A3 on most seeds (it clears $3/11$). The scheduler scorecard:

pre-registration	prediction	prior	outcome
CIFAR-10 2×2	P1 WaNet clean arm reverses with cosine; P2 BadNets reverses with the block	0.75, 0.70	both held (Table 45)
	P3 the annealed WaNet twin rises toward BadNets'	0.85	✓ ($0.9150 \rightarrow 0.9418$)
BPP / Input-Aware	P1–P2 BPP flips both ways; P3–P4 Input-Aware flips; P5 shipped twins $\geq$ 1.5 pt lower	0.65, 0.70, 0.55, 0.55, 0.75	P1, P2 and P5's BPP half held (twin gap $+2.79$ pt); P3, P4 and P5's Input-Aware half $\times$: untestable, SCHED2's K3 (a twin failing A4 is reset; failing again voids the attack) voided Input-Aware's clean arm (its ten-image twin fails A4 and its one-image reset diverges in training); its published cost moved in the predicted direction
CIFAR-100 (WaNet, Blended)	P1 shipped WaNet clean arm negative	0.75	$\times$ as registered (K3 voids the cell); ✓ as the declared deviation
	P2 cosine WaNet not significantly negative	0.70	✓
	P3 undefended twin moves ≥ 4.0 pt	0.60	$\times$ as registered (K3); ✓ as the declared deviation ($+6.01$)
	P4 fixed point, WaNet cells	0.65	$\times$ as registered (K3); ✓ as the declared deviation
	P5 Blended reverses when given the block	0.65	✓
	P6 shipped WaNet published cost negative	0.75	$\times$: untestable, K3 voids its test arm (attack installed in 1/5 seeds; on all five seeds it would have held, Table 44)
	P7 interaction invariant (< 0.5 pt)	0.55	$\times$: untestable, K3 voids its shipped side (on all five seeds it would have failed)
CIFAR-100 (BPP addendum)	P8 shipped BPP clean arm negative	0.70	✓
	P9 cosine BPP not negative	0.65	✓
	P10 twin moves ≥ 4.0 pt	0.60	✓ ($+6.75$)
	P11 four-cell fixed point (all three defenses)	0.60	$\times$ (fine-pruning)
SCHED4 (FT-SAM, SAU, i-BAU)	P12 FT-SAM flips as FT; defended twin within 1 pt of FT's	0.75	✓ (0.75 pt)
	P13 SAU flips; four-cell fixed point	0.65	✓ as registered (cosine WaNet $+2.88$ $[+1.93, +3.82]$; BadNets+block -0.58 $[-0.91, -0.25]$); strict two-sided reversal on BadNets only—on WaNet and BPP the never-annealed interval straddles zero
	P14 i-BAU moves in the predicted direction (shipped→swapped)	0.55	✓ (direction only; no sign change)
	P15 published WaNet cost rises with cosine for all three	0.70	✓

pre-registration	prediction	prior	outcome
SCHED5 (FT budget)	P16 defended-twin range shrinks monotonically with budget; $\geq$ 1.5 pt at 5 epochs of 10^{-2} and at 50 of 10^{-4}	0.70	$\times$ (monotone, but 0.71 and 1.39 pt: the fixed point forms faster than predicted)
	P17 a budget B^* exists above which the range is $<$ 1.0 pt; predicted 20–50 epochs at 10^{-2}	0.60	$\times$ as registered: B^* exists (✓) but is $\leq$5 epochs, the smallest budget run, against the predicted 20–50
	P18 clean-arm sign reversal at every 10^{-2} budget $\geq$ 10 epochs, all four intervals excluding zero; none at 10^{-4}	0.60	$\times$ as registered (at 50 epochs the BadNets-cosine interval is $+0.13$ $[-0.04,+0.29]$); both schedule contrasts reverse at every 10^{-2} budget $\geq$ 10, none at 10^{-4}
	P19 A3 fails for most poisoned arms at 5 epochs of 10^{-4}	0.70	✓ $(0/20)$
SCHED6 (VGG19-BN)	P20 clean arm negative shipped, positive cosine (FT, NAD), intervals excluding zero	0.70	✓ (FT -2.30 $[-2.72,-1.89]$ / $+2.09$ $[+1.59,+2.59]$)
	P21 twin swing $\geq$ 2.0 pt	0.65	✓ $(+3.12)$
	P22 FT/NAD defended twins within a third of the swing	0.60	$\times$ (1.72 pt)
SCHED7 (TSBD, CGD)	P23 TSBD gains on the shipped victim; cost $\geq$ 0 or straddles zero on cosine	0.70	✓ (-2.90 $[-3.49,-2.32]$, $+0.43$ $[-0.61,+1.47]$)
	P24 twin "gain" $\geq$ 1.0 pt shipped; cost $\geq$ 0 cosine	0.65	✓ (-1.11 $[-1.86,-0.35]$, $+2.59$ $[+1.37,+3.82]$)
	P25 defended twins within a third of the twin swing	0.50	✓ (0.92 pt against 0.93, a third of the five-seed swing $+2.78$)
	P26 CGD gains on the shipped victim; cost ≥ -1.0 or straddles zero on cosine	0.65	✓ (-3.69 $[-4.22,-3.16]$, -1.00 $[-2.10,+0.10]$)
	P27 CGD twin "gain" $\geq$ 1.0 pt shipped	0.60	✓ (-2.43 $[-3.00,-1.86]$)
SCHED9 (SIG)	P28 damaged annealed victim starts $\geq$ 5 pt below its twin, A0 holds	0.75	✓ $(+9.25)$
	P29 FT and NAD published cost negative, intervals excluding zero	0.80	✓ (-6.16 $[-6.40,-5.91]$)
	P30 defended victim within 1 pt of the defended twin (FT, NAD)	0.55	$\times$ (3.03 pt)
	P31 twin's clean-arm cost positive (FT, NAD)	0.70	$\times$ (FT $+0.07$ $[-0.12,+0.25]$; twins fail A4, Section S49)
	P32 DER's clip fires on the cell	0.80	✓
SCHED10 (tare column for the leaderboard)	P33 LF and Trojan: $\lvert g\rvert < 1.5$ pt, FT/NAD clean arm positive with intervals excluding zero	0.70	$\times$ as registered (LF FT $+0.11$ $[-0.03,+0.25]$; Trojan A4 $0/5$)
	P34 by mean tare, FT cheapest and ABL, I-BAU the two most expensive	0.60	$\times$ as registered over all seven attacks, where ANP and I-BAU swap (0.06 pt apart); ✓ on the A4-passing twins: FT $<$ NAD $<$ FP $<$ NC $<$ SAU $<$ ANP $<$ I-BAU (ABL off-scale, last)
	P35 prototype-referenced DER flips $\leq$ $4/28$ pairs, all near-ties	0.60	✓ $(0/28)$
	P36 on SIG and BPP at least 3 of the 5 added defenses share FT's tare sign	0.65	$\times$ (SIG $5/5$, BPP $1/5$ share fine-tuning's sign)

pre-registration	prediction	prior	outcome
SCHED11 (five further defenses, Section S52)	P37 CLP and NPD tare > 0 on all seven attacks; on WaNet and BPP intervals excluding zero in $\geq 3/4$	0.70	✓ $(3/4)$
	P38 MCR $\lvert\text{tare}\rvert < 0.5$ pt on annealed and < 1 pt on never-annealed victims	0.50	$\times$ as registered (WaNet -1.78)
	P39 RNP and D-BR destroy $\geq$ half their twins	0.60	$\times$ as registered (D-BR $35/50$, RNP $10/50$)
	P40 prototype-referenced DER flips $\leq 10\%$ of pairs, near-ties only	0.60	✓ on the seven registered attacks ($0/78$; $1/78$ on all nine)
	P41 FT cheapest by mean tare, CLP between FT and FP	0.55	$\times$ as registered: MCR $<$ FT $<$ NAD $<$ FP $<$ NC $<$ CLP $<$ SAU $<$ ANP $<$ I-BAU $<$ NPD $<$ RNP (ABL, D-BR off-scale, last)
SCHED12 (healthy start moved by training length, Section S50)	P42 the 30-epoch twin starts ≥ 1 pt below the 100-epoch twin	0.65	✓ $(4.41$ pt)
	P43 after FT and NAD the defended twins' range $\leq$ a third of the undefended range, slope ≤ 0.35; FP slope ≥ 0.5	0.60	$\times$ as registered: slopes FT $+0.32$, NAD $+0.31$, FP $+0.16$; ranges 1.78, 1.72, 1.24 pt against a third of 4.83, 1.61
	P44 FT's published cost ≥ 1 pt lower on the 30-epoch victim	0.60	✓ $(-1.96$ vs $+0.98)$

pre-registration	prediction	prior	outcome
SCHED13 (BadNets dose, Section S51)	P45 the starting gap g rises with the rate, Spearman $\rho \geq 0.8$ over the five rates	0.65	$\times$: untestable as registered: K20 voids the 0.5% and 1% poisoned arms its five-rate test needs; read there anyway, $\rho = 1.00$ over all five rates and 1.00 over the A0-passing ones
	P46 FT, NAD and FP paired contrast rises, $\rho \geq 0.8$; the 20% paired $\geq 2\times$ the 1%	0.60	$\times$ as registered: the rise ✓ (1.00, 1.00, 1.00); ratio clause untestable (K20 voided the 1% arm), counted as failed
	P47 at $\leq 1\%$ FT and NAD published cost within 0.5 pt of the tare	0.55	$\times$: untestable: K20 voided every rate $\leq 1\%$, counted as failed (A0 $0/5$, $0/5$)
SCHED14 (swapped cells to eleven seeds)	P48 eleven-seed pooled tare inside the five-seed interval, same sign	0.75	✓ (W-cos $+0.78$ $[+0.70,+0.87]$ vs $+0.82$ $[+0.72,+0.92]$; B-step -2.08 $[-2.40,-1.76]$ vs -2.03 $[-2.28,-1.77]$)
	P49 interval half-width shrinks $\geq 30\%$; $\geq 95\%$ of cells keep the sign	0.60	$\times$ as registered (half-width W-cos 0.10→0.09, 15% narrower; B-step 0.26→0.32, 22% wider; agreement $131/132$)
SCHED15 (FTrojanNN and Refool, Section S52)	P50 both twins pass A4 in $\geq 4/5$ seeds	0.65	✓ (FTrojanNN $5/5$, Refool $5/5$)
	P51 FT, FP and NAD tare positive with intervals excluding zero; published cost positive	0.70	✓ ($6/6$ intervals)
	P52 prototype-referenced DER flips $\leq 10\%$ of pairs, near-ties only	0.60	✓ ($0/28$; $1/78$ with the SCHED11 rows)
SCHED16 (Blended and LF dose, Section S51)	P53 A0 holds at every rate (1, 5, 10, 20%) on both attacks	0.75	$\times$ (void by K20: Blended none; LF 1%)
	P54 the starting gap rises with the rate, $\rho \geq 0.8$, each attack	0.60	$\times$ (1.00, 0.50)
	P55 FT and NAD published cost falls with the rate, $\rho \leq -0.8$, in ≥ 3 of the 4 (attack, defense) pairs	0.45	$\times$ ($0/4$)
	P56 FT and NAD paired contrast rises, $\rho \geq 0.8$, in ≥ 3 of 4	0.55	✓ ($4/4$)

pre-registration	prediction	prior	outcome
SCHED17 (BackdoorBox, Section S33)	P57 undefended twin ≥ 1 pt higher under A than under N	0.85	✓ (8.61 pt)
	P58 FineTuning (lr 0.001) tare negative under N, interval excluding zero	0.75	✓ (-6.10)
	P59 FineTuning (lr 0.001) tare positive under A, interval excluding zero	0.50	× ($+0.01$)
SCHED18 (three more BackdoorBox defenses)	P60 FineTuning at lr 0.01: tare negative under N and positive under A, both intervals excluding zero	0.75	× (-7.12, -0.01)
	P61 NAD tare higher under A than under N, seed-paired interval excluding zero	0.65	✓ ($+8.03$ [$+6.54$, $+9.53$])
	P62 NAD tare negative under N, interval excluding zero	0.55	✓ (-7.23)
	P63 Pruning, no training phase: tare ≥ 0 under both schedules	0.65	✓ ($+3.44$, $+1.69$)
SCHED19 (BackdoorBox BadNets)	P64 undefended twin ≥ 1 pt higher under A than under N	0.85	✓ (8.94 pt)
	P65 FineTuning at lr 0.01: tare negative under N and positive under A, both intervals excluding zero	0.75	× (-7.10, $+0.05$)
	P66 FineTuning (lr 0.001) tare negative under N, interval excluding zero	0.70	✓ (-5.91)
	P67 NAD tare higher under A than under N, seed-paired interval excluding zero	0.65	✓ ($+8.49$ [$+6.55$, $+10.43$])
	P68 Pruning, no training phase: tare ≥ 0 under both schedules	0.60	✓ ($+3.46$, $+1.80$)

pre-registration	prediction	prior	outcome
SCHED20 (zero-image twins, Section S34)	P69 zero-minus-ten undefended accuracy $\lvert\Delta\rvert \leq 0.3$ pt with interval covering zero, all three cells	0.85	✓ (max $\lvert\Delta\rvert$ 0.24)
	P70 $\geq$ 8 of 9 (cell, defense) tare differences with interval covering zero and $\lvert\Delta\rvert \leq 0.5$ pt	0.80	✓ $(8/9)$
	P71 all 15 zero-image twins pass A4	0.85	✓ $(15/15)$
SCHED21 (crossover at 1%/5%, Section S35)	P72 annealed victim starts $\geq$ 1 pt higher in $\geq 4/5$ seeds, every (attack, rate)	0.85	× $(2/4)$
	P73 WaNet: FT/FP/NAD published cost negative under the shipped block and higher under the swapped one, interval excluding zero, $\geq 5/6$	0.75	× $(3/6)$
	P74 BadNets: the same	0.75	× $(3/6)$
	P75 WaNet at 1% installs (A0 $\geq 3/5$) under both schedules	0.60	×
SCHED22 (ANP deployer-side rule, Section S36; analysis only)	P80 BadNets tare under rule D positive, interval excluding zero	0.70	✓ $(+6.44)$
	P81 rule D and rule O BadNets tares within one point	0.60	× (D − O $+3.95$)
	P82 WaNet tare of the same sign under both rules	0.60	✓ $(+1.78, +6.79)$
SCHED23 (eight-defense crossover, Section S37)	P76 ANP tare on WaNet positive under both schedules	0.60	✓
	P77 ABL destroys most twins under both schedules on WaNet and BPP	0.80	✓ (WaNet never $7/11$; WaNet cosine $7/11$; BPP never $3/5$; BPP cosine $5/5$)
	P78 NC fires on $\leq$ half the cosine WaNet twins and costs them a positive amount where it fires	0.65	× $(6/11; +3.31)$
	P79 the DER ordering of the eight on WaNet changes in a (fine-tuning-family, other) pair	0.60	✓ (FP/I-BAU, FP/SAU)

A "P7-type" invariance claim would need an equivalence bound; at five seeds we report it as a non-rejection (Section S8).

S44 THE TWIN FIXED POINT, ALL CELLS

Figure 11 plots all 162 twin–defense pairs of the six admissible cells.

S45 CROSSOVER AND EXTENSION TABLES

Table 44 gives the CIFAR-100 cells, Table 45 the CIFAR-10 2×2 crossover and Table 46 the BPP/Input-Aware extension.

Table 44: CIFAR-100 / PreAct-ResNet18, 5 seeds per cell. Clean-arm cost per defense, pooled published cost, undefended twin accuracy, and A3 (residual ASR ≤ 0.05) per defense. The shipped WaNet victim installs the attack in only $1/5$ seeds (ASR 0.767 and up), so by rule K3 (registered before any CIFAR-100 run; its disposition, a timestamped amendment written after seed 0 and before seeds 1–4, voids the poisoned arm and reports the clean arm as a deviation, Section S43) its poisoned-arm quantities are void. For the record, over all five seeds ignoring A0 the published cost is -5.24 $[-7.48, -2.99]$ pt, the interaction -1.37 $[-4.29, +1.56]$ pt and the victim base 63.12; evaluated on these, P6 would have held and P7 would have failed. Its clean arm does not depend on installation and is reported. Twins: ten images for WaNet, one image for BPP and Blended (Section S46).

attack	victim's scheduler	twin	fine-tuning	fine-pruning	NAD	clean arm (pooled)	published (pooled)	victim base	twin base	public victim	A4	A3 (ft/fp/nad)
WaNet	step [100, . . .] (shipped)	10 img	-4.54 $[-5.74, -3.34]$	-2.95 $[-4.41, -1.50]$	-4.11 $[-5.17, -3.04]$	-3.87 $[-5.07, -2.66]$	void (A0 1/5)	void	65.38	64.05	5/5	1/0/1
WaNet	cosine (swapped)	10 img	$+0.34$ $[-0.15, +0.83]$	$+4.41$ $[+3.68, +5.15]$	$+0.45$ $[-0.10, +1.00]$	$+1.73$ $[+1.43, +2.04]$	-0.62 $[-2.02, +0.78]^{\ddagger}$	68.53	71.39	—	5/5	3/0/2
Blended	cosine (shipped)	1 img	$+0.46$ $[+0.25, +0.68]$	$+7.55$ $[+6.93, +8.17]$	$+0.28$ $[+0.07, +0.48]$	$+2.76$ $[+2.50, +3.03]$	$+3.54$ $[+3.24, +3.84]^{\ddagger}$	69.47	70.87	69.28	5/5	0/0/0
Blended	step [100, . . .] (swapped)	1 img	-6.00 $[-8.20, -3.80]$	-3.97 $[-6.70, -1.25]$	-5.72 $[-7.92, -3.51]$	-5.23 $[-7.60, -2.86]$	-5.07 $[-7.46, -2.68]^{\ddagger}$	62.30	63.36	—	5/5	0/0/0
BPP	step [100, . . .] (shipped)	1 img (5%)	-4.89 $[-6.65, -3.13]$	-3.28 $[-5.54, -1.01]$	-4.62 $[-6.46, -2.79]$	-4.26 $[-6.20, -2.32]$	-3.96 $[-4.77, -3.14]$	65.01	65.02	63.96	5/5	3/2/5
BPP	cosine (swapped)	1 img (5%)	$+0.54$ $[+0.12, +0.97]$	$+6.20$ $[+5.23, +7.16]$	$+0.62$ $[+0.21, +1.02]$	$+2.45$ $[+1.95, +2.96]$	$+2.45$ $[+2.22, +2.67]$	71.51	71.77	—	5/5	5/0/5

Table 45: The 2×2 scheduler crossover on CIFAR-10 / PreAct-ResNet18: clean-arm cost c_{clean} (pt, positive = the defense cost accuracy on a model with no backdoor), t-interval over seeds. Only the victim's scheduler keys differ between a shipped row and its swapped partner. "twin base": the undefended twin's accuracy; "sign agree": (seed $\times$ defense) cells whose cost has the predicted sign (the one dissent is BadNets + cosine, fine-tuning, -0.03 pt). Every twin passes A4; A3 is in Table 26 and the WaNet victim's A0 in Section 4.2.

attack	victim's `lr_scheduler`	seeds	fine-tuning	fine-pruning	NAD	pooled	twin base	sign agree
WaNet	`MultiStepLR` (shipped)	11	-1.99 $[-2.30, -1.68]$	-1.68 $[-2.12, -1.24]$	-1.85 $[-2.14, -1.56]$	-1.84 $[-2.18, -1.50]$	91.50	33/33
WaNet	`CosineAnnealingLR` (swapped)	11	$+0.51$ $[+0.43, +0.59]$	$+0.86$ $[+0.74, +0.97]$	$+0.99$ $[+0.86, +1.12]$	$+0.78$ $[+0.70, +0.87]$	94.18	33/33
BadNets	`CosineAnnealingLR` (shipped)	11	$+0.13$ $[+0.06, +0.21]$	$+1.05$ $[+0.84, +1.25]$	$+0.33$ $[+0.24, +0.42]$	$+0.50$ $[+0.42, +0.59]$	93.71	32/33
BadNets	`MultiStepLR` (swapped)	11	-2.29 $[-2.61, -1.96]$	-1.95 $[-2.36, -1.53]$	-2.00 $[-2.30, -1.71]$	-2.08 $[-2.40, -1.76]$	91.09	33/33

Table 46: Extension to BPP and Input-Aware, CIFAR-10 / PreAct-ResNet18, 5 seeds. Clean-arm cost per defense and pooled (per-seed mean over the three defenses, t-interval over seeds), and pooled published cost. A4: twin admissible (ASR ≤ 0.05); the attack installs in every seed (A0 $5/5$ throughout); ‡ marks a published number on which fewer than half the cells clear A3. BPP's twin is one poisoned image (2×10^{-5}) because ten images already install a 5–7% ASR on an annealed BPP victim (Section S46). Input-Aware's twin fails A4 at ten images and diverges at one; its clean-arm row is shown, not claimed.

attack	victim's scheduler	twin	fine-tuning	fine-pruning	NAD	clean arm (pooled)	published (pooled)	victim base	twin base	A4	A3
BPP	step $[100, \ldots]$ (shipped)	1 img	-2.18 $[-2.86, -1.49]$	-1.87 $[-2.48, -1.25]$	-2.00 $[-2.62, -1.39]$	-2.01 $[-2.63, -1.40]$	-2.01 $[-2.42, -1.60]$	91.10	91.31	$5/5$	$10/15$
BPP	cosine (swapped)	1 img	$+0.55$ $[+0.30, +0.80]$	$+0.97$ $[+0.76, +1.17]$	$+1.11$ $[+0.92, +1.30]$	$+0.87$ $[+0.69, +1.06]$	$+1.28$ $[+1.16, +1.41]$	94.08	94.10	$5/5$	$10/15$
Input-Aware	step $[100, \ldots]$ (shipped)	10 img	-2.14 $[-2.56, -1.73]$	-1.93 $[-2.37, -1.49]$	-1.93 $[-2.48, -1.38]$	-2.00 $[-2.46, -1.55]$	-1.84 $[-2.53, -1.15]^{\ddagger}$	91.21	90.96	$0/5$	$6/15$
Input-Aware	step $[30, 60, 90]$ (swapped)	10 img	$+0.42$ $[+0.25, +0.59]$	$+1.07$ $[+0.77, +1.36]$	$+0.49$ $[+0.27, +0.70]$	$+0.66$ $[+0.54, +0.77]$	$+0.61$ $[+0.48, +0.74]^{\ddagger}$	93.09	93.38	$0/5$	$5/15$

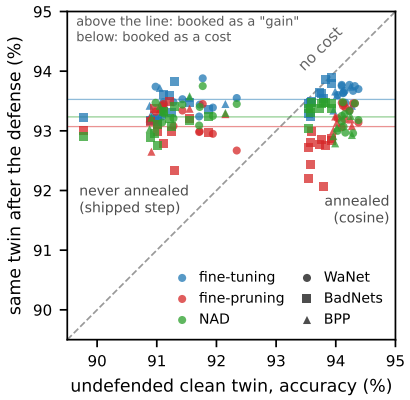


Figure 11: The per-defense fixed point. Each point is one clean twin from one of the six admissible cells (WaNet, BadNets, BPP $\times$ two schedulers; 162 twin–defense pairs): its accuracy before the defense (x) and after (y). Never-annealed twins start near 91, annealed ones near 94; every defense returns both to the same 93–93.5. Above the diagonal the benchmark books a "gain", below it a cost. The scheduler decides which side a victim starts on.

S46 The twin convention

A twin is the attack's own pipeline run at a poisoning rate too small to install the attack. Ten images (2×10^{-4}) is clean for WaNet and BadNets (max twin ASR 0.038 and below). On BPP with the cosine schedule ten images install a 0.045–0.068 ASR in $4/5$ seeds (A4 $1/5$); the pre-registration's disposition for this case was to reset to one image, which passes A4 on both schedules ($5/5$, $5/5$); Blended on CIFAR-100 uses the one-image twin from the start, following BPP and gives readings within 0.15 pt of the ten-image ones ($+0.58$ [$+0.24$,$+0.91$]/$+0.94$ [$+0.75$,$+1.13$]/$+0.96$ [$+0.66$,$+1.27$] vs. $+0.55$ [$+0.30$,$+0.80$]/$+0.97$ [$+0.76$,$+1.17$]/$+1.11$ [$+0.92$,$+1.30$]). On Input-Aware ten images carry 0.17–0.46 ASR and one image makes the generator's joint training diverge (NaN loss at epoch 26–28 in all ten runs); no further rate was tried. On BPP's shipped schedule the ten-image twin passes A4 ($5/5$) and is the twin Table 4 reports; Tables 36 and 46 use the one-image reset for both BPP rows; on shipped BPP, where both twins exist, they read within 0.16 pt of each other for FT, FP, NAD and I-BAU and 0.53 pt apart for SAU.

What the matched twin buys. The recipe match, not variance reduction. Across installations the poisoned and twin arms' pre-defense accuracies correlate at $r = -0.27$ (BackdoorBench BadNets, $n = 11$), -0.38 (WaNet, $n = 11$) and $+0.19$ (our CIFAR-10 headline grid, $n = 8$), and after each defense on the two BackdoorBench grids at -0.74 to $+0.43$, so pairing by seed does not in general narrow the paired contrast's interval against an unpaired one; what the twin controls is the recipe, schedule and data order the defense starts from, which is where the tare comes from.

S47 The fixed point as a function of budget

Pre-registered as SCHED5 (Section S43): FT with the benchmark's configuration except `--epochs` and `--lr`, on all 40 models of the four CIFAR-10 cells (480 runs), twelve budgets plus the benchmark's own. Figure 12 and Table 47 give the defended clean-twin level per cell and the four-cell range. One fact about the benchmark's trainer, found after the sweep had run and recorded in the pre-registration as a post-hoc note: it builds `CosineAnnealingLR` with $T_{\max} = 100$ regardless of `--epochs` (`utils/aggregate_block/train_settings_generate.py`, line 67; no FT configuration sets `cos_t_max`), so a budget of E epochs is the first E epochs of a 100-epoch cosine, not a shorter complete one—the learning rate at the last epoch is 0.00996 at five, 0.00914 at twenty and 0.00516 at fifty. Two regimes: at 10^{-2} the range is already 0.71 pt after five epochs, before any decay, and 0.27 pt after twenty (K8, range ≥ 1.5 at 20 epochs, did not fire); at 10^{-4} the twins keep their order and the range is 1.39 pt after fifty epochs. Of the shipped WaNet twin's $+2.01$ pt rise over the full schedule, $+1.52$ pt arrives in the first five epochs at 10^{-2} and

$+0.49$ pt over the ninety-five that contain the decay; the annealed twin first drops -1.04 pt in those five epochs and climbs back to -0.56 pt by the end. So the level the defenses share is set mostly by the fine-tuner's own regime (learning rate 10^{-2}, batch 256 against the victim's 128, the 5% clean subset) and only in its last half-point by the decay: "the defense completes the anneal" (our first phrasing) is about a quarter of the story, and the main text now says "return a low start to a level of their own." In the first regime the clean-arm cost is the start-to-fixed-point distance of Section 3.2 at every budget; in the second it is small for every cell (shipped WaNet twin -0.26 $[-0.81,+0.28]$ pt at five epochs, -0.95 $[-1.47,-0.43]$ pt at fifty) and the poisoned arm is not cleaned (A3 $0/20$ at five epochs of 10^{-4}), so the published cost then reports neither the backdoor nor the start. Two of the four predictions failed on their thresholds, both in the same direction: the fixed point forms faster than we predicted (P16, P17's B^* of 5 epochs against a predicted 20–50), and P18's four-interval criterion fails at 50 epochs of 10^{-2} because the BadNets-cosine cell, whose cost is near zero by construction, straddles it ($+0.13$ $[-0.04,+0.29]$); the two schedule contrasts themselves reverse at every 10^{-2} budget from ten epochs up and at none of the 10^{-4} budgets.

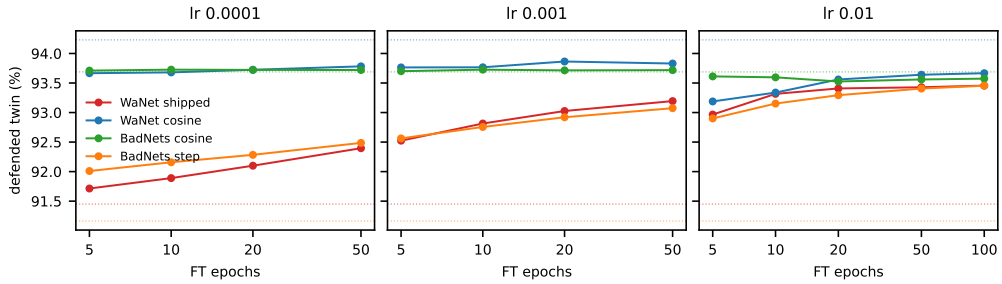


Figure 12: Defended clean-twin accuracy against FT epochs, one panel per learning rate, one line per cell; dotted lines are the undefended twins. At 10^{-2} the four lines meet within a point from five epochs on; at 10^{-4} they keep their starting order through fifty.

Table 47: Defended clean-twin accuracy (%, mean of 5 seeds) per cell and budget, and the four-cell range (pt).

lr	epochs	WaNet shipped	WaNet cosine	BadNets cosine	BadNets step	range
0.0001	5	91.71	93.67	93.71	92.01	2.00
0.0001	10	91.89	93.68	93.73	92.16	1.84
0.0001	20	92.10	93.72	93.72	92.28	1.62
0.0001	50	92.40	93.78	93.72	92.49	1.39
0.001	5	92.52	93.76	93.70	92.56	1.24
0.001	10	92.81	93.77	93.73	92.76	1.01
0.001	20	93.03	93.86	93.71	92.92	0.94
0.001	50	93.19	93.83	93.72	93.07	0.76
0.01	5	92.97	93.19	93.61	92.90	0.71
0.01	10	93.32	93.34	93.60	93.15	0.44
0.01	20	93.41	93.56	93.53	93.29	0.27
0.01	50	93.43	93.64	93.56	93.41	0.23
0.01	100	93.46	93.67	93.57	93.45	0.21

S48 The WaNet pair on VGG19-BN

Pre-registered as SCHED6 (Section S43): the WaNet crossover of Section 4.2 with `--model vgg19_bn`, five seeds per cell, FT, FP and NAD (Table 48). The defenses run with AMP off on this backbone (the benchmark's AMP path aborts with a floating-point exception on VGG19-BN on our machine, so all thirty defense calls of the first launch produced nothing; the attack side trains without AMP by default), identically under both schedules, recorded as a timestamped amendment

before any VGG19-BN defense result existed. Undefended twins: 89.41% shipped, 92.53% cosine, a swing of +3.12 pt (the leaderboard's VGG19-BN gap is +2.6 pt); A4 5/5 and 5/5, A0 4/5 and 5/5. The reversal is as clean as on PreAct-ResNet18: clean-arm cost −2.30 [−2.72,−1.89] pt shipped against +2.09 [+1.59,+2.59] pt cosine under FT, −2.18 [−2.58,−1.78] against +1.86 [+1.69,+2.03] under FP, −1.92 [−2.40,−1.44] against +2.92 [+2.26,+3.58] under NAD, every interval excluding zero (P20 ✓, P21 ✓). The fixed point is not: the defended twins land at 91.72% from the shipped start and 90.44% from the cosine start under FT (largest FT/NAD gap 1.72 pt, against a third of the swing), so P22 fails—on this backbone the benchmark's fine-tuning takes a well-trained clean model down by two points and an under-trained one up by two, and the two do not meet. The published cost on the poisoned arm is a start-point story with wide intervals: the shipped victims start between 71.8 and 87.9% (WaNet at 10% damages VGG19-BN unevenly across seeds), and FT books −8.44 [−20.58,+3.69] pt against −0.76 [−5.02,+3.49] pt on the cosine victims; A3 is 2/4 shipped and 4/5 cosine under FT. What transfers across backbones is therefore the sign of the clean-arm cost and its dependence on the three keys; the common level that all defenses return to is a property of PreAct-ResNet18 on these datasets, and Section 3.2 should be read with that scope.

Table 48: VGG19-BN / CIFAR-10 WaNet cells, 5 seeds, t-intervals; costs in pt, negative is a "gain".

victim's scheduler	defense	published cost	clean-arm cost	defended victim	defended twin	A3	seeds
step (shipped)	FT	−8.44 [−20.58,+3.69]	−2.30 [−2.72,−1.89]	91.56	91.72	2/4	5
	FP	−8.35 [−20.27,+3.57]	−2.18 [−2.58,−1.78]	91.46	91.60	2/4	5
	NAD	−7.98 [−20.04,+4.08]	−1.92 [−2.40,−1.44]	91.09	91.33	2/4	5
cosine (swapped)	FT	−0.76 [−5.02,+3.49]	+2.09 [+1.59,+2.59]	89.92	90.44	4/5	5
	FP	−1.05 [−5.13,+3.04]	+1.86 [+1.69,+2.03]	90.20	90.67	3/5	5
	NAD	−0.09 [−3.94,+3.76]	+2.92 [+2.26,+3.58]	89.24	89.61	5/5	5

S49 A SECOND CAUSE OF A LOW START: SIG AT 10%

Pre-registered as SCHED9 (Section S43) after the VGG19-BN cell, to test the leaderboard's reading that the clip fires wherever the victim starts below the defense's level, whatever made it low. SIG is a clean-label attack that leaves every scheduler key alone—its victim anneals under the prototype's cosine schedule—and at the benchmark's 10% rate its poisoning damages the victim outright: five seeds start at 84.58% (84.4–84.7, ASR 0.980–0.989, A0 5/5) against a same-recipe twin at 93.83%, a gap of +9.25 points with no scheduler involved. Fine-tuning books the rise as a gain (published cost −6.16 [−6.40,−5.91]; NAD −5.22 [−5.38,−5.06], fine-pruning −5.89 [−6.25,−5.53]), so DER's clip fires on this cell (P32), while on the twin the same defenses cost +0.07 [−0.12,+0.25], +0.31 [+0.04,+0.59] and +1.23 [+0.79,+1.66]. The defended arms do not meet: 90.74% against 93.76% under fine-tuning (3.03 points; NAD 3.72), so P30 fails as registered—the damage is erased only in part, and the fixed-point accounting of Section 3.2 holds here for the sign of the published cost and not for its size. Two things this cell cannot say. Its twins fail A4 (1/5; twin ASR up to 0.213) for a reason the protocol's rule does not distinguish from installation: the SIG signal itself moves a clean model's predictions (the leaderboard's own 0.1% SIG victim has an ASR of 41%), a one-image reset would not change that, and we made none; the twin's clean-arm numbers are therefore descriptive, and P31 (fine-tuning's clean-arm cost positive with its interval excluding zero) fails on that arm. And A3 clears in 4/5 of the poisoned arms under fine-tuning.

Table 49: SIG at 10% on CIFAR-10 / PreAct-ResNet18, five seeds, t-intervals; costs in pt, negative is a "gain"; the twin columns are descriptive (A4 1/5).

defense	published cost	clean-arm cost	defended victim	defended twin	A3	seeds
FT	−6.16 [−6.40,−5.91]	+0.07 [−0.12,+0.25]	90.74	93.76	4/5	5
FP	−5.89 [−6.25,−5.53]	+1.23 [+0.79,+1.66]	90.47	92.61	4/5	5
NAD	−5.22 [−5.38,−5.06]	+0.31 [+0.04,+0.59]	89.80	93.52	4/5	5

S50 A HEALTHY START MOVED BY TRAINING LENGTH

Pre-registered as SCHED12 (Section S43) to test whether the level of Section 3.2 belongs to the defense when a healthy, fully annealed victim's start is moved by something other than the scheduler block: BadNets under the prototype's cosine schedule trained for 30 and 60 epochs (the 100-epoch cells are seeds 0–2 of the shipped grid), three installation seeds, poisoned arm and twin, then FT, NAD and FP at their shipped configurations (Table 50). The start moved (4.41 pt between the 30- and 100-epoch twins, P42 ✓) and every victim installed (A0 3/3, 3/3, 3/3). The three defenses compress that spread to 1.78, 1.72 and 1.24 pt (slopes of defended on undefended twin over the nine twin arms: +0.32, +0.31, +0.16), each to a level of its own (FT 92.18–93.63, NAD 91.97–93.44, FP 91.87–92.37%), so the accounting holds approximately and not as an identity on a healthy victim: P43 fails as registered (the range criterion by 1.78 against a third of 4.83, and FP's slope is below, not above, 0.5). What the block-free manipulation reproduces is the sign: FT's published cost on the 30-epoch victim is −1.96 against +0.98 on the 100-epoch one (P44 ✓), so a fully annealed victim whose start sits below the defense's level is booked as a gain and clipped by DER exactly as a never-annealed one is.

Table 50: SCHED12: BadNets under the prototype cosine schedule at three training lengths, 3 seeds each. Undefended twin and victim (A0), the tare per defense (t-intervals over seeds) and the defended twin's level.

epochs	seeds	twin	victim (A0)	FT tare	NAD tare	FP tare	FT level	NAD level	FP level
30	3	89.35	89.54 (3/3)	−2.84 [−3.56,−2.11]	−2.62 [−3.60,−1.64]	−2.52 [−3.65,−1.40]	92.18	91.97	91.87
60	3	92.92	91.55 (3/3)	−0.33 [−0.61,−0.06]	+0.03 [−0.22,+0.28]	+0.12 [−0.45,+0.68]	93.25	92.89	92.80
100	3	93.76	91.62 (3/3)	+0.12 [−0.18,+0.43]	+0.32 [−0.04,+0.68]	+1.39 [+0.59,+2.19]	93.63	93.44	92.37

S51 MORE POISON, A LOWER START, A CHEAPER-LOOKING REMOVAL

Pre-registered as SCHED13 (Section S43): BadNets on CIFAR-10 / PreAct-ResNet18 under the prototype's cosine schedule at poisoning rates 0.5%, 1%, 5%, 10%, 20% (the 10% cells are seeds 0–4 of the shipped grid; the seed-0 victims at 1, 5 and 20%, with their FT and NAD runs, predate the registration, which says so), five installation seeds per rate, the eight defenses at their shipped configurations (32 GPU-hours). The twin does not depend on the rate, so one twin per seed serves every rate and the tare is the same number in every row by construction. K20 voids the poisoned arm at 0.5%, 1%, where the attack does not install (A0 0/5 and 0/5); P47, which needed those rates, and P46's ratio clause, which needed 1%, are untestable and counted as failed. The starting gap rises over all five rates (−0.09 at 0.5% to +3.58 at 20%). P45 was registered over the five rates, but K20 reports only A0 and the twin at a voided rate, so at 0.5%, 1% its registered test is untestable and counted as failed; read there anyway, $\rho = 1.00$ over all five rates, 1.00 over the A0-passing ones. On the A0-passing rates the paired contrast rises for FT, NAD and FP (Spearman 1.00, 1.00, 1.00) and the published cost falls: for all eight defenses between 5% and 20% (8/8, the utility floor applied; Spearman −1.00, −1.00, −1.00 for FT, NAD, FP). Per seed, over the 4 seeds whose poisoned victim installs at both rates, fine-tuning's published cost changes by −0.82 [−1.43,−0.21] while the starting gap grows +2.33 [+1.50,+3.15] (the g column of Table 51 averages all five seeds). SCHED13's file had named the opposite, a published cost rising with the rate, as half of its falsifiable signature; that half failed, and the fall was read off this table before SCHED16 registered it for Blended and LF (Table 52). The identity of Section 3.2 (published cost = tare + paired contrast − starting gap) does not fix its sign; the data do: the start drops by as much as the starting gap grows (the twin is fixed) while the defended victim's level drops by less, by as much as the paired contrast rises, so the defense recovers part of what the extra poison took and books the recovery as a smaller cost. A defense whose level did not depend on the start at all (Section 3.2's fixed point) would have cut the published cost one for one; fine-tuning's changes by about a third as much (−0.82 against a starting-gap rise of +2.33 per seed): its defended level follows the start part of the way.

The registered replication on Blended and LF (SCHED16). Registered before any run (Section S43) to test the fall on two further annealed CIFAR-10 attacks whose 10% installs reach 0.99,

Table 51: SCHED13: BadNets dose–response, five seeds per rate. A0 per rate, the starting gap g (twin minus victim), and for FT, NAD and FP the tare (one twin per seed for every rate), the published cost and the paired contrast (A0-passing seeds; — where K20 voids the poisoned arm).

rate	A0	g	FT tare	NAD tare	FP tare	FT published	NAD published	FP published	FT paired	NAD paired	FP paired
0.5%	0/5	−0.09	+0.11	+0.26	+1.11	—	—	—	—	—	—
1%	0/5	+0.24	+0.11	+0.26	+1.11	—	—	—	—	—	—
5%	4/5	+1.17	+0.11	+0.26	+1.11	+1.36	+2.26	+0.07	+2.48	+3.22	+0.32
10%	5/5	+2.06	+0.11	+0.26	+1.11	+0.94	+2.15	−0.31	+2.89	+3.96	+0.65
20%	5/5	+3.58	+0.11	+0.26	+1.11	+0.52	+1.48	−0.52	+3.99	+4.80	+1.95

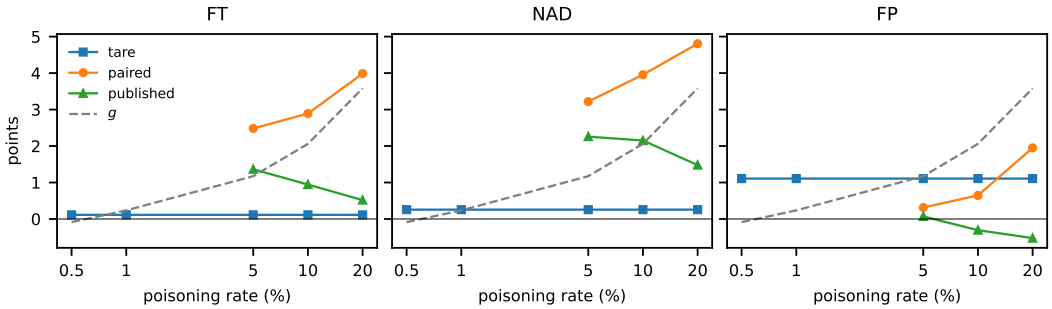


Figure 13: SCHED13, the rows of Table 51 against the poisoning rate (log scale): the tare is flat by construction, the starting gap g and the paired contrast rise, the published cost falls.

Blended and LF, at 1, 5, 10 and 20% with five seeds each, the eight defenses and the same twins (Table 52). A0 holds at every rate on Blended and fails at 1% on LF (0/5), so P53 fails. The start gap rises with the rate on Blended (Spearman 1.00) and not monotonically on LF (0.50: its 5% and 10% seed means, +0.562 and +0.560, differ below the printed precision; P54 fails as registered). The paired contrast rises for FT and NAD on both attacks (4/4 pairs; P56 holds), and the published cost rises with it: 0/4 of the four (attack, defense) pairs fall (Spearman +1.00, +1.00, +1.00, +1.00), so P55, the primary prediction, fails in the opposite direction. Per seed between the lowest and highest installing rate, fine-tuning's published cost changes by +1.06 [+0.67,+1.44] on Blended and +0.96 [+0.67,+1.24] on LF while the start gap grows by +1.02 [+0.78,+1.27] and +0.50 [+0.17,+0.82], against +2.33 [+1.50,+3.15] on BadNets. The identity of Section 3.2 reads the three attacks together—with the twin fixed, the published cost falls where the victim's start drops by more than its defended level (BadNets) and rises where it drops by less (Blended, LF)—but that reading is after the fact: SCHED13's fall is a BadNets observation, not a replicated one.

Table 52: SCHED16: Blended and LF dose–response, five seeds per rate (the 10% rows are the shipped seeds 0–4), laid out as Table 51; — where A0 fails in at least three of five seeds (K20).

attack	rate	A0	g	FT tare	NAD tare	FP tare	FT published	NAD published	FP published	FT paired	NAD paired	FP paired
Blended	1%	5/5	−0.01	+0.22	+0.40	+1.10	+0.16	+0.41	+1.06	−0.07	−0.01	−0.06
Blended	5%	5/5	+0.23	+0.22	+0.40	+1.10	+0.48	+1.14	+0.72	+0.49	+0.96	−0.15
Blended	10%	5/5	+0.51	+0.22	+0.40	+1.10	+0.52	+1.54	+1.00	+0.81	+1.65	+0.41
Blended	20%	5/5	+1.01	+0.22	+0.40	+1.10	+1.22	+2.29	+1.04	+2.01	+2.89	+0.95
LF	1%	0/5	+0.33	+0.11	+0.22	+1.08	—	—	—	—	—	—
LF	5%	5/5	+0.56	+0.11	+0.22	+1.08	+0.31	+1.01	+0.77	+0.76	+1.35	+0.25
LF	10%	5/5	+0.56	+0.11	+0.22	+1.08	+0.87	+1.62	+0.91	+1.32	+1.96	+0.39
LF	20%	5/5	+1.06	+0.11	+0.22	+1.08	+1.26	+2.02	+0.95	+2.21	+2.86	+0.93

S52 THE CIFAR-10 LEADERBOARD AT 10% WITH ITS TARE COLUMN: PUBLISHED COSTS, DER AND SEEDS

Pre-registered as SCHED10 (Section S43) and extended by SCHED11 (five further defenses) and SCHED15 (FTrojanNN, Refool). For every attack of the benchmark's CIFAR-10 / 10% roster that admits a never-poisoned twin under this protocol—BadNets, Blended, LF, FTrojanNN, Refool, Trojan, SIG, WaNet, BPP—and the eight defenses of Table 29, Table 53 prints the tare (the clean-arm cost on the twin, unconditional) and Table 54 the published cost on the victim (A0-gated; ‡ where fewer than half the cells clear A3). Twins are the attack's own pipeline at pratio 0.0002 (ten images; one image for Blended, whose ten-image twin fails A4). Seed by seed on our grids (FT, FP, NAD, NC, I-BAU; Table 54 prints the cell means), DER's clip fires in $58/65$ never-annealed, $22/25$ SIG (damaged by its own poison) and $13/195$ healthy seed-cells, all fine-pruning's (BadNets $11/11$, Refool $2/5$). Two attacks' twins fail A4 and their tare columns are descriptive: SIG's signal moves a clean model's predictions (A4 $1/5$, Section S49), and Trojan's twins answer its trigger (A4 $0/5$); whether its optimised trigger fires on any same-recipe model or ten images install it we did not test. Attacks without an admissible twin under this protocol are not printed: Input-Aware (its ten-image twin fails A4 too, its one-image twin diverges, and only FT, FP and NAD were run on it, Section S45), LIRA and Blind (no per-sample poisoning rate), SSBA and LC (the benchmark does not ship the encoded or adversarial data they need, and we download nothing), and Poison Ink and CTRL (a trigger generator or a self-supervised pipeline would have to be trained first). LF and Trojan are new here, and FTrojanNN and Refool (SCHED15; the checkout ships their frequency trigger and reflection images; neither is on the public leaderboard) were added last; Trojan's trigger optimisation needs a pretrained clean model the benchmark does not ship, and we used the seed-0 BadNets twin (the prototype recipe plus ten poisoned images) for it. The table's sign pattern is the paper's accounting read across a roster; the sign counts and the ordering below use only the 7 attacks whose twins pass A4 (BadNets, Blended, LF, FTrojanNN, Refool, WaNet, BPP), the Trojan, SIG rows being descriptive, and leave ABL out: it destroys $50/60$ twins, so under the registered kill rule K16 (a defended twin that loses more than 20 points is a failed run) its row is a failure record, printed and marked ♮ but outside every sign count and the ordering. Outside ABL's row no run in Table 4 is excluded as a failed run (a defended twin that lost more than 20 points) or by A3's utility floor (Section S3), which keeps 7 of ABL's poisoned-arm runs out of its A3 count. Over all nine attacks, among these eight defenses K16 fired once outside ABL (SIG seed 4 SAU (defended twin 54.26%); that cell is printed over the remaining seeds, marked ♭), and the utility floor of A3 (Section S3) excludes 2 poisoned-arm runs outside ABL (Trojan/SAU seed 3, SIG/SAU seed 2) from A3 and from the paired contrast, and 14 of ABL's from its A3 count only. On the 5 admissible annealed victims every tare is positive ($35/35$ cells); on Refool, whose victim starts 1.67 points below its twin, fine-pruning's published cost ($+0.08$ $[-0.25,+0.40]$) already sits below its tare ($+0.91$ $[+0.62,+1.21]$). On the two never-annealed victims FT, FP and NAD "gain" in $6/6$ cells and ANP charges the twin in $2/2$, every interval excluding zero; SAU, which re-runs a full schedule but lands at a level of its own (Section 4.3), and I-BAU, which has none, sit at zero in $4/4$ of their cells; Neural Cleanse's detector fires on $3/11$ WaNet and $0/5$ BPP twins, and where it fires its repair, the benchmark's own 100-epoch fine-tune, "gains" -1.30 as fine-tuning does. This pattern was read off the table after P36 failed on BPP, not predicted. Ordering the defenses by mean tare over the admissible attacks gives FT $<$ NAD $<$ FP $<$ NC $<$ SAU $<$ ANP $<$ I-BAU (ABL off-scale, last) (over all nine attacks, FT $<$ NAD $<$ FP $<$ NC $<$ SAU $<$ ANP $<$ I-BAU $<$ ABL). A prototype-referenced DER (Section 4.5; the same seed's BadNets twin as reference) shifts every defense's mean DER down by 0.5 to 1.1 points and re-ranks $0/28$ of the defense pairs over all nine attacks (Table 57; on the public twelve-defense panel of Section S41 it re-ranks near-ties only): the reference changes attribution, not order. Scorecard, judged on the seven attacks registered: P33 $\times$ (LF's fine-tuning interval touches zero, $+0.11$ $[-0.03,+0.25]$; Trojan's twins fail A4), P34 $\times$ (over all seven registered attacks ANP and I-BAU swap; on their A4-passing twins ✓), P35 ✓, P36 $\times$ (on SIG $5/5$ of the five added defenses share fine-tuning's sign; on BPP $1/5$, SAU alone, the pattern above). SCHED15, for the two attacks added last: P50 ✓ (A4 FTrojanNN $5/5$, Refool $5/5$), P51 ✓ ($6/6$ fine-tuning-family tares positive with intervals excluding zero), P52 ✓ ($0/28$).

Table 53: The tare per (attack, defense) cell: clean-arm cost on the never-poisoned twin (pt; negative is a "gain"), mean over seeds with its t-interval; the victim's schedule, the undefended twin and A4 beside it. $^\flat$: one K16 run excluded (SIG seed 4 SAU (defended twin 54.26%)), $n-1$ seeds; $^\natural$: ABL, a K16 failure record ($50/60$ twins destroyed), outside every sign count and the ordering. †: A2 holds at the policy level only (Table 29, Appendix S19); ANP's threshold is chosen by test-set attack success within an accuracy budget, which on a twin selects on false-positive noise. Neural Cleanse's cell is printed as the fraction of twins on which its detector fired | the cost conditional on firing (the others are returned unchanged). Intervals are per-cell and nominal: of the 63 cells printed with one, 56 exclude zero at 0.05 and 20 at the Bonferroni level $0.05/K$, $K = 63$ (44 and 15 of the 49 admissible-attack cells); the sign counts in the text are read from the cell means, not from these intervals.

attack	schedule	seeds	twin (%)	A4	FT	FP	NAD	NC†	ANP†	ABL†♮	I-BAU	SAU
BadNets	cosine	11	93.71	11/11	+0.13 ±0.07	+1.05 ±0.20	+0.33 ±0.09	5/11 | +2.71	+2.49 ±2.10	+33.20 ±4.06	+5.70 ±1.71	+2.70 ±0.39
Blended	cosine	8	93.84	8/8	+0.20 ±0.11	+1.12 ±0.23	+0.36 ±0.15	3/8 | +3.06	+3.86 ±2.86	+27.83 ±11.98	+4.95 ±1.02	+3.06 ±1.76
LF	cosine	5	93.92	5/5	+0.11 ±0.14	+1.08 ±0.46	+0.22 ±0.18	1/5 | +3.22	+3.25 ±3.08	+31.28 ±15.75	+5.98 ±2.26	+2.83 ±0.80
FTrojanNN	cosine	5	93.84	5/5	+0.24 ±0.15	+1.04 ±0.36	+0.47 ±0.25	3/5 | +2.86	+1.43 ±1.01	+32.36 ±12.73	+6.06 ±2.37	+2.92 ±0.51
Refool	cosine	5	93.74	5/5	+0.16 ±0.15	+0.91 ±0.30	+0.36 ±0.16	1/5 | +2.79	+4.72 ±3.62	+35.91 ±11.28	+5.66 ±1.68	+3.71 ±1.95
Trojan	cosine	5	93.90	0/5	+0.20 ±0.14	+1.01 ±0.24	+0.23 ±0.15	4/5 | +2.58	+6.53 ±1.00	+37.50 ±11.55	+4.82 ±1.51	+2.56 ±0.40
SIG	cosine (damaged)	5	93.83	1/5	+0.07 ±0.18	+1.23 ±0.44	+0.31 ±0.28	1/5 | +3.98	+7.27 ±1.75	+34.59 ±10.97	+6.64 ±1.42	+3.91 ±4.61$^\flat$
WaNet	never	11	91.50	11/11	−1.99 ±0.31	−1.68 ±0.44	−1.85 ±0.29	3/11 | −1.30	+1.78 ±1.26	+27.77 ±6.49	+0.23 ±0.53	−0.15 ±0.62
BPP	never	5	91.48	5/5	−2.07 ±0.75	−1.82 ±1.07	−1.84 ±0.64	0/5 | —	+4.28 ±2.48	+29.83 ±12.18	+0.71 ±1.67	−0.60 ±0.96

Table 54: The published cost per (attack, defense) cell on the poisoned victim (pt; A0 seeds), with the undefended victim and A0 beside it; ‡: fewer than half the cells clear A3 (residual attack success ≤ 0.05 and the utility floor of Section S3).

attack	schedule	seeds	A0	victim (%)	FT	FP	NAD	NC†	ANP†	ABL†♮	I-BAU	SAU
BadNets	cosine	11	11/11	91.64	+1.01	−0.22	+2.10	+1.11	+6.50	+9.39	+3.85	+1.39
Blended	cosine	8	8/8	93.41	+0.62‡	+0.97‡	+1.58‡	+1.28‡	+7.03	+20.31	+4.10	+4.12
LF	cosine	5	5/5	93.36	+0.87‡	+0.91‡	+1.62‡	+2.43	+6.78	+24.40‡	+4.81‡	+2.52
FTrojanNN	cosine	5	5/5	93.46	+0.70‡	+0.81‡	+1.36‡	+1.90	+6.84	+16.51	+4.36‡	+2.19
Refool	cosine	5	5/5	92.06	+0.18‡	+0.08‡	+0.68‡	+0.00‡	+6.33	+39.14‡	+4.25‡	+1.76‡
Trojan	cosine	5	5/5	93.30	+0.86‡	+0.86‡	+1.79‡	+1.46	+7.29	+20.85	+4.01	+8.18
SIG	cosine (damaged)	5	5/5	84.58	−6.16	−5.89	−5.22	−2.21‡	+5.30	+32.41‡	−1.60	+5.68
WaNet	never	11	8/11	90.05	−3.14	−3.15	−2.92	−1.45	+5.15	+20.72‡	−0.96	−1.68
BPP	never	5	5/5	91.10	−2.22	−1.93‡	−1.88	−1.59	+6.42	+27.40‡	+0.40	−0.41

S53 TSBD AND CGD MEASURED IN THE WANET 2×2

Pre-registered as SCHED7 (Section S43). TSBD's released code was dropped into our BackdoorBench checkout unchanged except for device and record paths; its clean reference model is the same-seed BadNets-cosine twin's weights (no new model was trained). We used the paper's printed defaults ($n = 0.15$, $m = 0.7$, 20 fine-tuning epochs at 10^{-2}, 5% clean data). One disclosure: the released code sweeps the re-initialisation ratio over 11 values from 0.01 to 0.9 and never selects one, so our first harvest took the last row ($n = 0.9$), which destroys the network; we fixed the operating point to the printed default before anything entered the paper (timestamped in the pre-registration) and record the whole sweep in Table 59. The sign pattern of Table 58—shipped victim gains, cosine twin loses—holds at 9 of the 11 ratios; the shipped victim gains more than the cosine victim at all 11, and the cosine twin loses more than the shipped twin at all 11; at $n \geq 0.5$ the twins lose +2.9 pt or more. Stage 1 (clean unlearning, then zero re-initialisation) leaves the shipped victim at 78.0% and the shipped twin at 45.7% (cosine: 79.3% and 60.4%); stage 2, plain fine-tuning with a gradient-norm regulariser, is what carries all four to the defended values in Table 58. P23–P25: ✓, ✓, ✓ (P25 by 0.92 against 0.93 pt, a third of the five-seed swing—a near miss, not a margin); K11 did not fire. CGD (P26 ✓ on its interval clause—the cosine-victim point estimate sits on the −1.0 line—, P27 ✓) used its released two-step code: CLIP ViT-B/32 separation, then 10 epochs of fine-tuning on the retained subset; our victims are at the benchmark's 10% rate, the CGD paper's at 5%.

Table 55: SCHED11: the five further defenses that run unchanged on the same arms (Section S43; EP, BNP and NAB do not run in this checkout, `LOCAL_CHANGES.md`; D-ST failed under its registered override and is dropped, Section S43). Tare per cell as in Table 53; $^{\flat}$ marks a cell with K16 exclusions (RNP: $11/60$ twins lost more than 20 points, printed over the rest), $^{\natural}$ a failure record (D-BR destroys $45/60$ twins), † RNP's threshold is chosen by test-set attack success. RNP leaves $23/60$ twins within a point of the undefended one and cuts others by up to 74 points (median over all runs 8.4); NPD makes 3 never-poisoned twins answer the trigger (attack success BadNets 0.99, 0.71; FTrojanNN 0.64).

attack	schedule	seeds	twin (%)	A4	CLP	NPD	MCR	RNP†	D-BR$^{\natural}$
BadNets	cosine	11	93.71	11/11	$+0.74$ ± 0.42	$+2.64$ ± 0.30	-0.01 ± 0.07	$+1.76$ $\pm 3.00^{\flat}$ (n=7)	—
Blended	cosine	8	93.84	8/8	$+1.32$ ± 0.83	$+2.49$ ± 0.29	-0.04 ± 0.08	$+3.98$ $\pm 5.58^{\flat}$ (n=7)	—
LF	cosine	5	93.92	5/5	$+0.96$ ± 0.39	$+2.75$ ± 0.41	-0.01 ± 0.06	$+0.10$ $\pm 0.31^{\flat}$ (n=4)	—
FTrojanNN	cosine	5	93.84	5/5	$+0.73$ ± 0.37	$+2.71$ ± 0.53	$+0.05$ ± 0.09	$+4.01$ ± 7.44	—
Refool	cosine	5	93.74	5/5	$+1.15$ ± 0.82	$+2.51$ ± 0.51	$+0.01$ ± 0.09	$+0.52$ $\pm 1.13^{\flat}$ (n=4)	—
Trojan	cosine	5	93.90	0/5	$+1.83$ ± 1.14	$+2.95$ ± 0.69	$+0.01$ ± 0.14	$+13.01$ $\pm 13.50^{\flat}$ (n=4)	—
SIG	cosine (damaged)	5	93.83	1/5	$+0.47$ ± 0.22	$+2.51$ ± 0.68	$+0.04$ ± 0.09	$+1.83$ $\pm 5.54^{\flat}$ (n=4)	—
WaNet	never	11	91.50	11/11	$+0.96$ ± 0.34	$+8.06$ ± 3.88	-1.78 ± 0.28	$+12.38$ $\pm 2.89^{\flat}$ (n=10)	$+10.07$ ± 2.64 (n=10)
BPP	never	5	91.48	5/5	$+1.54$ ± 1.77	$+5.45$ ± 5.40	-1.83 ± 0.91	$+11.45$ $\pm 7.63^{\flat}$ (n=4)	$+8.96$ ± 3.28

Table 56: SCHED11: the published cost per cell on the poisoned victim for the same five defenses (A0 seeds; ‡: fewer than half the cells clear A3, which is most of them: CLP and NPD leave the backdoor in place on most attacks at their shipped configurations; among the admissible attacks MCR clears it on BadNets, WaNet, BPP and not on Blended, LF, FTrojanNN, Refool). The public leaderboard prints the same shape at 10% on PreAct-ResNet18: D-BR leaves 29.41 (BadNets), 14.77 (Blended), 11.23 (LF) and 12.03 (TrojanNN) clean accuracy, 61.70 on WaNet and 83.11 on BPP; ABL's published BadNets cost there is $+8.00$ against $+9.39$ $[+8.12,+10.65]$ here; RNP has no public CIFAR-10 cell.

attack	schedule	seeds	A0	victim (%)	CLP	NPD	MCR	RNP†	D-BR$^{\natural}$
BadNets	cosine	11	11/11	91.64	$+1.82^{\ddagger}$	$+3.46^{\ddagger}$	$+0.12$	$+57.70^{\ddagger}$	$+75.45^{\ddagger}$
Blended	cosine	8	8/8	93.41	$+1.19^{\ddagger}$	$+2.39^{\ddagger}$	$+0.40^{\ddagger}$	$+61.51^{\ddagger}$	$+80.43^{\ddagger}$
LF	cosine	5	5/5	93.36	$+1.85^{\ddagger}$	$+2.54$	$+0.25^{\ddagger}$	$+57.17^{\ddagger}$	$+78.96^{\ddagger}$
FTrojanNN	cosine	5	5/5	93.46	$+0.96$	$+2.45$	$+0.39^{\ddagger}$	$+34.55^{\ddagger}$	$+83.01^{\ddagger}$
Refool	cosine	5	5/5	92.06	$+0.78^{\ddagger}$	$+2.34^{\ddagger}$	$-0.35^{\ddagger}$	$+72.80^{\ddagger}$	$+72.00^{\ddagger}$
Trojan	cosine	5	5/5	93.30	$+0.81^{\ddagger}$	$+2.23^{\ddagger}$	$+0.40$	$+79.21^{\ddagger}$	$+81.73^{\ddagger}$
SIG	cosine (damaged)	5	5/5	84.58	$+0.87^{\ddagger}$	$-4.16^{\ddagger}$	-5.96	$+46.21^{\ddagger}$	$+67.75^{\ddagger}$
WaNet	never	11	8/11	90.05	$+2.34$	$+0.20^{\ddagger}$	-2.97	$+40.38^{\ddagger}$	$+15.52^{\ddagger}$
BPP	never	5	5/5	91.10	$+1.41$	$+17.79^{\ddagger}$	-1.77	$+35.61^{\ddagger}$	$+5.70^{\ddagger}$

S54 CITATION AUDIT

The full audit table (paper, arXiv id, venue, victim provenance, verbatim clean-accuracy claim, where it appears, whether load-bearing, whether it rests on the three cells), the search queries with hit counts, and the 11 journal-only papers that could not be obtained (`AUDIT.md`, §1D) are in the supplement.

The three claims in detail. The main text compresses the following.

Roster. We pulled the 252 papers citing either BackdoorBench paper (Semantic Scholar, deduplicated), read the 48 that propose a defense against installed backdoors and could be obtained in full, and asked of each: does it report WaNet, BPP or Input-Aware on a BackdoorBench victim, does it make a clean-accuracy claim, and does that claim survive removing those cells? Thirteen reuse the benchmark's victims—their "No Defense" rows reproduce the leaderboard to two decimals—and their baseline columns reproduce the artifact: FT, FP, NAD and NC "gaining" two points on WaNet and Input-Aware appears, unremarked, in the tables of at least seven.

Three claims rest on it. *TSBD* (Lin et al., 2024) opens its CIFAR-10 results with an average accuracy "higher than the original No Defense models," attributed to "removing the backdoor effect

Table 57: DER per defense, mean over the attacks above: the benchmark's clipped, victim-referenced form and the prototype-referenced form of Section 4.5 (reference: the same seed's BadNets twin). Published order SAU > ANP > I-BAU > NAD > FP > FT > NC > ABL; prototype order SAU > ANP > I-BAU > NAD > FP > FT > NC > ABL; with the SCHED11 defenses added, $1/78$ of the pairs re-rank (FT–NPD, 0.03 DER points apart).

DER (mean over the 9 attacks)	FT	FP	NAD	NC	ANP	ABL	I-BAU	SAU
published (clipped, victim-referenced)	82.1	84.3	86.3	79.6	95.4	75.4	93.9	95.8
prototype-referenced	81.6	83.8	85.7	78.8	94.3	74.3	93.0	95.0

Table 58: TSBD and CGD on the WaNet victims and their clean twins, CIFAR-10 / PreAct-ResNet18, t-intervals over seeds, A0-gated. Published cost $= Q(\theta_{\text{pois}}) - Q(D\theta_{\text{pois}})$; clean-arm cost likewise on the twin; negative is a "gain".

method	victim's scheduler	published cost	clean-arm cost	defended victim	defended twin	A3	seeds
TSBD	step (shipped)	-2.90 [-3.49,-2.32]	-1.11 [-1.86,-0.35]	92.94	92.56	4/4	5
TSBD	cosine (swapped)	$+0.43$ [-0.61,$+1.47$]	$+2.59$ [$+1.37$,$+3.82$]	92.06	91.64	5/5	5
CGD	step (shipped)	-3.69 [-4.22,-3.16]	-2.43 [-3.00,-1.86]	93.73	93.88	4/4	5
CGD	cosine (swapped)	-1.00 [-2.10,$+0.10$]	$+0.41$ [$+0.26$,$+0.57$]	93.49	93.82	5/5	5

with the least cost": 91.70% against 91.34% over eight attacks (+0.36 pt, no interval reported). The margin is carried by the two never-annealed cells (Input-Aware +2.39, WaNet +2.01 pt) and by SIG (+5.93 pt), a damaged annealed victim. Dropping the two never-annealed cells, the average is 91.21 against 91.46; subtracting from each of the two the 1.84 pt that the benchmark's fine-tuning family (FT, FP and NAD pooled) gains on the clean twin leaves 91.24 against 91.34; subtracting the victim's 2.64 pt scheduler deficit leaves 91.04. The sign reverses under each accounting. TSBD itself, run in the WaNet 2×2, confirms it (Section S53). *CGD* (Xu et al., 2025) reports that it "can identify and unlearn poisoned samples, improving CA over the No Defense baseline by 3.4% for WaNet and 3.6% for IAB"—the sign survives, the attribution does not. CGD fine-tunes the backdoored model on a CLIP-separated clean subset at 5% poisoning, a shorter schedule than the benchmark's; what transfers from our measurements is the victim's deficit (+2.1 pt at that rate, Table 3), not a value, and CGD itself, run in the same 2×2 (Section S53), gives the value: published cost -3.69 [-4.22,-3.16] pt on the shipped victim and -1.00 [-2.10,$+0.10$] pt on the cosine victim; the clean twin gains -2.43 [-3.00,-1.86] pt under the shipped schedule and loses $+0.41$ [$+0.26$,$+0.57$] pt under cosine, so 66% of the shipped-victim gain is what a never-poisoned model gets from the same fine-tuning. The sentence attributes all of them to unlearning. *PDB* (Wei et al., 2024) is a training-time defense: it does not fine-tune the victim but trains its own model, so the mechanism there is not a defense completing the anneal but a comparison against a No-Defense baseline that the attack's schedule left under-annealed. Its Tiny-ImageNet/ViT-B/16 table has a positive average accuracy change that is produced entirely by WaNet (+11.9) and BPP (+10.2); the other four attacks are all negative, and the same two cells carry the table's highest DER. The two gains match the +11.3 pt leaderboard gap on that backbone (Table 3); but Tiny-ImageNet is the once-firing contrast on which we ran no swap and on which VGG19-BN reverses (Section 4.4), so for PDB the attribution to the scheduler is correlational, causal only for TSBD and CGD. Two further papers (Min et al., 2024; 2023) have load-bearing no-cost statements that these cells buffer without reversing, and two more (Wei et al., 2023; 2025) report DER values that the same cells lift through the clip. *None* of the 48, nor either benchmark paper, notes that the three victims are under-trained; the closest is FST's appendix, which lowers WaNet's learning rate "since a large initial learning rate would make the attack collapse" without connecting the resulting gap to the fine-tuning gains it then reports. One author coded all 48 papers; no second coder or inter-rater agreement was measured. The audit's criterion, written before coding (`AUDIT.md`), also defines a loose count that adds the load-bearing claims the three cells buffer without reversing (PDB, ULRL and FST) to the strict 2 the text uses.

We are not saying these methods do not work. Every one of the three has other evidence. We are saying that a specific sentence in each depends on these cells. For TSBD and CGD we have

Table 59: TSBD across the released code's re-initialisation sweep: mean cost (pt) per cell at the final fine-tuning epoch. The paper's printed default is the operating point of Table 58.

reinit. ratio n	shipped victim	shipped twin	cosine victim	cosine twin
0.01	-3.09	-1.93	$+0.19$	$+1.43$
0.05	-3.16	-1.78	$+0.08$	$+1.68$
0.1	-3.02	-1.52	$+0.34$	$+2.08$
0.15 (default)	-2.90	-1.11	$+0.43$	$+2.59$
0.2	-2.77	-0.74	$+0.48$	$+3.06$
0.25	-2.60	-0.24	$+0.88$	$+3.57$
0.3	-2.49	$+0.15$	$+0.88$	$+4.31$
0.4	-1.53	$+1.24$	$+1.66$	$+5.88$
0.5	-0.60	$+2.91$	$+2.92$	$+10.10$
0.7	$+2.69$	$+10.40$	$+6.26$	$+20.23$
0.9	$+18.30$	$+31.57$	$+27.33$	$+37.46$

shown causally that the victims in those cells start two to three points low because of the scheduler; for TSBD and CGD the anneal is measured on the method itself (Section S53). For PDB even the victim's deficit is attributed only on the leaderboard. And a benchmark whose cost metric cannot go negative let those sentences be written.

S55 Changes from the registered version

The abstract registered on OpenReview (submission 16702, 2026-09-18) is replaced by the one above. Apart from the exceptions listed below (two from changed data, three typed), no number printed in the registered version changed value: every macro the registered draft inputs from `supplement/build/` is regenerated by the same builders from the same data, and only `data/scheduler_crossover_cells.json` and `data/tare_z_cifar.csv` differ from the registered supplement. The first exception is the NAD row of the TARE-Z calibration (Table 35), which the registered version had computed on the retired lr-0.01 teacher; it is recomputed on the corrected teacher of Appendix S30 (eight new NAD runs on the surrogate models), moving the pooled error from -0.01 $[-0.14, +0.11]$ paired and -0.02 $[-0.51, +0.47]$ DiD to -0.02 $[-0.17,+0.12]$ and -0.03 $[-0.48,+0.42]$, and with it the per-defense Kendall τ_a of Appendix S28 ($+0.714 \to +0.857$ over all eight, $+0.714 \to +0.905$ without ANP). The second is the two swapped crossover cells, which SCHED14 extended from five to eleven seeds after registration: their pooled tares move from $+0.82$ $[+0.72,+0.92]$ to $+0.78$ $[+0.70,+0.87]$ (WaNet-cosine) and -2.03 $[-2.28,-1.77]$ to -2.08 $[-2.40,-1.76]$ (BadNets-step), the cell agreement from $95/96$ to $131/132$, and every number of Table 2 and Appendix S21 built on them is recomputed on eleven seeds (the six-defense Table 36 and the fixed-point ranges of Section 3.2 stay on the five registered seeds). The three typed corrections: ABL's clean-arm lower bound in Table 30 ($+29.2 \to +29.1$, a rounding fix), the base-model drift of the language-model row ($1.0 \to 1.08$ points, now generated), and the CIFAR-10 twin range of Section 3.3 (now over every admissible twin). What changed is scope and emphasis: the main text now leads with the scheduler block as the cause (Section 4.2, Table 2) and carries the leaderboard with its tare column as its deliverable (Section 4.5, Table 4); the estimand section states the identity between the two estimators and prints the tare and the paired contrast rather than correcting for an unidentified removal share (Section 3.1); the language-model row moved to Table 7. Added: the public-leaderboard fixed point (Section 4.4); the block's replications on BPP, six defenses, CIFAR-100 and VGG19-BN, the budget sweep, and the SIG cell (Section 4.3, Appendices S31–S49); TSBD and CGD re-run with their released code and the citation audit (Section 4.4, Appendices S53, S54); the paired and g columns of Table 25; the fixed-point accounting of Section 3.2. Removed from the abstract, each still printed where stated: "one extra run suffices, -0.01 points" (Section 4.7, now -0.02 after the NAD correction above, with its domain of validity); the 354% ABL share (ABL is off-scale under a policy-level A2; Table 30); "the residue TARE isolates does not move" (rewritten as the start gap returned, Section 4.2); the Qwen $+11.46 \to +1.37$ headline (Table 7, with its independence interval in the caption); the four-pair crossover band and the WaNet "none survives" (Appendix S19); the clean-data budget sweep's $3.79 \to 0.28$ (Section 3.3, Appendix S16); the TOFU contrast (Appendix S23) and the Llama row (Table 8; one installation, three defense

seeds); and the sentence attributing the never-annealed victims' recovery to the defense "completing the anneal", which the budget sweep does not support (Appendix S47). Added last: the tare column for the 10% leaderboard (Section 4.5, Appendix S52; SCHED10, LF and Trojan new, five defenses added on SIG and BPP; 35 GPU-hours), then SCHED11 (five further defenses), SCHED12 (a healthy start moved by training length), SCHED13 (BadNets dose), SCHED14 (eleven seeds in the swapped cells), SCHED15 (FTrojanNN and Refool; with SCHED10 and SCHED11, 118 GPU-hours for Section 4.5), SCHED16 (the dose replication on Blended and LF, Appendix S51) and SCHED17–SCHED19 (the block in BackdoorBox, Appendix S33); the per-defense intervals, fixed-point slopes, 2×2 pooled tares and TSBD/CGD point values of the interim abstract that replaced the registered one on OpenReview are now in Figure 6, Section 4.4, Table 2 and Section 4.4. The scheduler-pipeline tables in Appendices S31–S54 gate every poisoned-arm quantity, the A3 fraction included, on A0; the registered tables condition published and DiD quantities on A3 and print A3 over all seeds; the two conventions are stated where they meet and the clean-arm columns, which neither gates, agree.